UNIVERSITÉ DE MONTRÉAL

GÉNÉRATION DE SCÉNARIOS PAR QUANTIFICATION OPTIMALE EN DIMENSION ÉLEVÉE

SIMON PROULX
DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(MATHÉMATIQUES APPLIQUÉES)
JUIN 2014

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

GÉNÉRATION DE SCÉNARIOS PAR QUANTIFICATION OPTIMALE EN DIMENSION ÉLEVÉE

présenté par : PROULX Simon
en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

M. AUDET Charles, Ph.D., président
M. LABIB Richard, Ph.D., membre et directeur de recherche
M. LE DIGABEL Sébastien, Ph.D., membre

REMERCIEMENTS

Je tiens d'abord à remercier mon directeur de recherche Richard Labib d'avoir accepté de diriger ma maîtrise et de m'avoir proposé un sujet de recherche adapté à mon grand intérêt pour le domaine des probabilités. Il a su tirer le maximum de mes capacités en posant des questions pertinentes qui m'ont aidé à rédiger un travail mieux réfléchi. Sa grande disponibilité, ses précieux conseils et son soutient continu ont grandement facilité le chemin vers l'accomplissement de ma maîtrise.

J'aimerais également remercier François Soumis de nous avoir suggéré un sujet de recherche passionnant et de s'être chargé de créer des liens avec l'industrie privée afin que le projet porte sur un sujet réellement appliqué. Ses conseils et son expertise dans le domaine de la recherche opérationnelle m'ont été très bénéfiques.

Je souhaite remercier Guy Desaulniers et François Soumis pour leur aide financière qui m'a permis de me consacrer pleinement à la recherche au cours de mes études.

Merci à Antoine Legrain et Lê Nguyên Hoang, qui m'ont beaucoup aidé respectivement au début et à la fin de ma maîtrise.

Merci également à KRONOS, particulièrement à Ahmed Baba Hadji et Beyime Tachefine, qui ont pris la peine de nous rencontrer et qui ont travaillé fort pour nous fournir les données sans lesquelles mon projet ne bénéficierait pas de la même crédibilité.

RÉSUMÉ

L'optimisation stochastique est une branche de la recherche opérationnelle qui traite de problèmes d'optimisation impliquant des phénomènes aléatoires. Le programme mathématique, défini par une fonction objectif et des contraintes, contient alors une ou plusieurs variables aléatoires et est appelé programme stochastique. Avant de procéder à sa résolution, on doit d'abord modéliser le vecteur aléatoire apparaissant dans le problème. Or, il n'est généralement pas possible de résoudre un programme stochastique pour lequel le support des variables aléatoires est infini, tel que les distributions continues. On doit donc discrétiser la loi probabiliste sur un ensemble fini de réalisations, appelées scénarios, auxquelles on assigne une probabilité. Leur quantité est choisie en fonction du temps de convergence numérique et de la finesse de l'approximation désirés. L'objectif principal de ce mémoire est de développer des méthodes de génération de scénarios permettant d'obtenir des solutions près de l'optimalité pour les programmes stochastiques comprenant des vecteurs aléatoires en haute dimension.

Il existe plusieurs méthodes de génération de scénarios, dont la plus courante est l'échantillonnage pur et consiste à tirer les scénarios au hasard à partir de la distribution estimée du vecteur aléatoire. Cependant, les scénarios obtenus par échantillonnage pur sont rarement ceux qui représentent le mieux la loi de probabilités. Dans ce mémoire, nous justifions à l'aide de résultats théoriques et expérimentaux que les scénarios devraient plutôt être générés par la méthode de quantification optimale. Nous montrons ensuite que lorsque le nombre de données tend vers l'infini, les problèmes de k-médianes et k-moyennes sont équivalents à la quantification optimale avec les normes L_1 et L_2 respectivement. Les techniques développées pour générer les scénarios sont donc inspirées d'algorithmes de partitionnement de données.

Il n'est pas toujours possible d'estimer avec confiance la distribution d'un vecteur aléatoire à partir d'un ensemble de données. Le cas où la distribution est connue (ou estimable) est donc traité séparément de celui où elle ne l'est pas. Lorsque la distribution est connue et que le problème ne contient qu'une seule variable aléatoire, nous utilisons l'algorithme de Lloyd qui nous permet d'atteindre le minimum global des problèmes de k-moyennes et k-médianes continus. Dans le cas multidimensionnel, nous choisirons plutôt la méthode de quantification vectorielle par apprentissage compétitif (QVAC). La quantification optimale suggère l'utilisation de la distance induite par la norme L_1, puisqu'elle permet d'établir une borne supérieure sur l'erreur de discrétisation du programme stochastique. Afin de quantifier le vecteur aléatoire avec la norme L_1, nous adaptons le paramètre de saut de la QVAC, qui

est généralement utilisé avec la distance euclidienne. Nous trouvons cependant que la borne supérieure sur l'erreur de discrétisation peut être beaucoup plus grande que l'erreur elle-même. On en déduit que la norme L_2 peut également être utilisée pour générer les scénarios et offre une plus grande couverture des événements extrêmes. Lorsque la distribution n'est pas connue, nous utilisons les algorithmes d'échange des centres et de Lloyd (k-médianes et k-moyennes) qui permettent de générer les scénarios directement à partir des données.

Dans le dernier chapitre, on analyse entre autres les effets du nombre de scénarios, de la norme utilisée, de la variance et de la dimension du vecteur aléatoire sur nos méthodes de génération de scénarios. On observe sans surprises qu'il est particulièrement difficile d'obtenir des solutions de qualité lorsque la dimension est élevée. Trois méthodes sont donc proposées pour réduire la dimension effective du problème, dont l'analyse par composantes principales et les copules. Parmi celles-ci, on constate cependant que seule l'analyse par composantes principales permet de réduire les coûts de l'optimisation stochastique en dimension élevée.

Les scénarios sont testés sur le problème du vendeur de journaux, où la demande suit une loi log-normale ainsi que sur une application réelle à partir d'un ensemble de données historiques lié à la confection d'horaire de personnels. Les solutions de l'optimisation stochastique à partir des méthodes de génération de scénarios proposées ont été comparées à celles obtenues par échantillonnage pur et par l'optimisation déterministe. Pour le problème du vendeur de journaux avec distribution de probabilités connue, des gains substantiels de nos méthodes sont observés pour la quasi-totalité des instances étudiées. Lorsque la distribution n'est pas connue, nos méthodes induisent des erreurs de discrétisation moins de 340 fois plus petites que celles de l'optimisation déterministe avec 100 scénarios. Les erreurs obtenues par les algorithmes d'échange des centres et de Lloyd sont similaires, mais ce dernier reste généralement plus pratique à cause de sa simplicité et sa rapidité d'exécution. Dans le cas du problème de confection d'horaires, nous utilisons toutefois l'algorithme d'échange des centres puisqu'il permet de générer des scénarios de la demande en nombres entiers. Malgré la dimension élevée du problème et les faibles variances, nos méthodes de génération de scénarios permettent tout de même d'obtenir des gains modestes par rapport à l'optimisation déterministe.

ABSTRACT

Stochastic optimization is a branch of operations research that deals with optimization problems involving random processes. The mathematical program defined by the objective function and the constraints then contains one or several random variables and is called stochastic program. Prior to its resolution, we must first model the random vector appearing in the problem. However, it is generally not possible to solve a stochastic program for which the support of the random variables is infinite, such as continuous distributions. Therefore, we must discretize the probabilistic law over a finite set of events, called scenarios, to which we assign probabilities. Their quantity is chosen according to the desired convergence time and discretization precision. The main objective of this paper is to develop scenario generation methods that lead to near optimal solutions of the stochastic program containing high dimensional random vectors.

There are several methods for generating scenarios, amongst which the most common is pure sampling and consists in randomly selecting scenarios from the estimated distribution. However, the scenarios obtained by pure sampling are rarely those which represent the law of probability best. In this paper, we justify using experimental and theoretical results that the scenarios should rather be generated dy optimal quantization. We then show that when the data set is infinite, the clustering analysis k-means and k-medians problems are equivalent to optimal quantization with L_1 and L_2 norm respectively. As a result, the techniques developed for generating scenarios are inspired by clustering analysis algorithms.

It is not always possible to confidently estimate the distribution of a random vector from data. The cases where the probability distribution is known (or can be estimated) and unknown are thus treated separately. In the event where the probability law is known and the problem includes a single random variable, we use Lloyd's algorithm, which converges to the global minimum of the continuous k-medians and k-means problems. In the multivariate cases, we will rather chose the competitive learning vector quantization (CLVQ) method. Optimal quantization suggests the use of the L_1-norm, since it allows us to establish an upper bound on the stochastic program discretization error. In order to quantify the random vector with the L_1-norm, we adapt the CLVQ step parameter, which is ordinarily used with euclidean distance. However, we find that the upper bound on the discretization error may be much larger than the error itself. Hence, we deduce that the L_2-norm may also be used to generate scenarios and provides greater coverage of extreme events. When the distribution is

unknown, we use the swapping centers and Lloyd (k-medians and k-means) algorithms that allow direct scenario generation from the data.

In the last chapter, we analyze the effects of the number of scenarios, norm, variance and dimension of the random vector on our scenario generation methods. As expected, we observe that it is particularly difficult to obtain quality solutions when the dimension is high. Three methods are then proposed to reduce the effective dimensionality of the problem, including principal components analysis and copulas. Amongst these, it is noted that only principal components analysis reduces the costs of high-dimensional stochastic optimization.

Our scenarios are tested on a virtual news vendor problem, where demand follows a log-normal distribution, and on a real data set for employee scheduling. The stochastic optimization solutions obtained by our methods are compared to those of pure sampling and deterministic optimization. For the news vendor problem with known probability distribution, substantial gains of our methods are observed for almost all instances studied. When the distribution is unknown, our methods induce costs less than 340 times that of deterministic optimization with 100 scenarios. The discretization errors obtained by the swapping center algorithm and Lloyd's method are similar, but the latter is most appealing due to its simplicity and execution speed. Nonetheless, for the employee scheduling problem, we still use the swapping algorithm since it allows us to generate integer scenarios of demand. Despite the high dimensionality of the problem and low variances, our scenario generation methods allow modest profits compared to deterministic optimization.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

ACP	Analyse par Composantes Principales
EC	Échange des Centres
EP	Échantillonnage Pur
OD	Optimisation Déterministe
PD	Programmation Déterministe
PS	Programmation Stochastique
QVAC	Quantification Vectorielle par Apprentissage Compétitif

CHAPITRE 1

INTRODUCTION

La recherche opérationnelle est un domaine d'études combinant les mathématiques à l'informatique, et qui se sert d'outils tels que l'optimisation, la modélisation et l'analyse stochastique dans le but de prendre les meilleures décisions pour les problèmes de gestion. Ses applications sont variées et touchent entre autres aux problèmes de réseaux de transport, de confection d'horaires, de finances, de localisation optimale et d'approvisionnement.

Étant donné leur complexité, les problèmes de recherche opérationnelle font presque toujours appel à l'optimisation. On s'intéressera aux cas où un phénomène aléatoire influence les décisions optimales ; on parle alors d'optimisation stochastique. On utilisera le terme «optimisation déterministe» pour désigner les cas où aucune variable aléatoire n'est en jeu, ou encore lorsque cette dernière n'est réduite qu'à un seul événement (sa moyenne par exemple) avec probabilité 1. L'avantage de l'optimisation stochastique est qu'elle permet normalement de réduire les coûts ou d'accroître les gains du problème considéré par rapport à son homologue déterministe. En revanche, elle demande plus d'efforts, que ce soit par la génération de scénarios ou la résolution du programme stochastique, en plus d'accroître le temps d'exécution lié à l'optimisation.

Le problème d'optimisation est défini par une fonction objectif, un ensemble de contraintes et des variables de décisions. Lorsque celui-ci contient également des variables aléatoires, il est appelé programme stochastique (PS). Ce dernier ne peut généralement pas être résolu lorsque le support de la distribution est infini (i.e. lorsque la variable aléatoire peut prendre une infinité de réalisations), sauf pour quelques cas particuliers tels que le problème du vendeur de journaux (*newsvendor problem*) qui peuvent être résolus analytiquement (Birge et Louveaux (2011)). Les algorithmes numériques comme la méthode *L-shaped* ne peuvent quant à eux considérer qu'un ensemble fini d'événements lors de la résolution. Il faut donc quantifier la distribution sur un ensemble fini de réalisations nommées scénarios, dont la qualité influence directement la solution optimale. Ceux-ci approximent la distribution et doivent être choisis de manière à la représenter le plus fidèlement possible.

La génération de scénarios est donc une étape cruciale de l'optimisation stochastique. Diverses méthodes de génération de scénarios existent et leur performance dépend des carac-

téristiques du problème étudié telles que la dimension, la distribution de la variable aléatoire, le nombre d'étapes et la complexité du programme stochastique. La dimension représente notamment un défi de taille lorsqu'elle est élevée, car plusieurs problèmes deviennent alors intraitables ; une propriété surnommée «le fléau de la dimensionnalité» par Richard Bellman. Certains programmes stochastiques comprennent plusieurs étapes, où une décision est prise après chaque observation d'un vecteur aléatoire. On s'intéressera aux problèmes stochastiques possédant deux étapes de décision, qui sont courants en pratique. Les décisions de première étape sont donc prises avant l'unique réalisation du vecteur aléatoire et celles de deuxième étape (appelées recours) sont prises après.

Plusieurs problèmes de gestion sont régis par un grand nombre de phénomènes probabilistes. On n'a qu'à penser à la gestion d'un portefeuille comportant diverses actions soumises aux aléas de la bourse ou encore à la distribution de l'électricité à travers le réseau, où la production de chaque centrale ainsi que la demande des clients sont aléatoires. Les problèmes de transport collectif, aérien et de collecte des déchets sont souvent influencés par le climat, le trafic, etc. Une entreprise désirant ouvrir un entrepôt peut aussi devoir choisir une localisation en fonction de facteurs externes tels que la demande de la clientèle et le trafic routier. En effet, il une multitude d'exemples de problèmes influencés par une grande quantité de variables aléatoires et pour lesquels l'optimisation stochastique s'avère utile. Le but ultime de ce mémoire est de développer une méthode performante permettant de générer des scénarios en dimension élevée pour les programmes stochastiques à deux étapes. Celle-ci sera testée sur le problème de confection d'horaires des employés dans un supermarché, où le nombre d'employés requis à chaque période de la journée est considéré comme aléatoire.

Plan du mémoire : Nous commençons d'abord par présenter une revue de littérature sur la génération de scénarios. Nous présentons et justifions ensuite le choix de la méthode utilisée et démontrons le lien étroit entre celle-ci et le partitionnement de données. Certaines méthodes de partitionnement de données sont présentées pour finalement être testées et analysées dans le dernier chapitre sur le problème du vendeur de journaux et sur une application liée à la confection d'horaire de personnel pour une entreprise.

CHAPITRE 2

REVUE DE LITTÉRATURE

La méthode de génération de scénarios la plus simple est sans doute l'échantillonnage pur, qui consiste à échantillonner une distribution supposément connue et d'utiliser les valeurs obtenues comme scénarios. Cependant, le coût et les décisions optimales de l'optimisation stochastique obtenus par cette procédure risquent de varier fortement d'un échantillon à l'autre. Pour contrer ce problème, il est possible d'utiliser des techniques de réduction de la variance tels que l'échantillonnage antithétique, stratifié ou préférentiel (Lemieux (2009)). Infanger (1992) utilise l'échantillonnage préférentiel et obtient des coûts assez près de l'optimalité avec de petits échantillons. Plutôt que de générer les scénarios dès le départ, Higle et Sen (1991) utilisent une méthode appelée décomposition stochastique. Celle-ci génère des réalisations de la variable aléatoire à chaque itération de l'algorithme, basé sur une combinaison de la méthode *L-shaped* (Van Slyke et Wets (1969)) et des quasi-gradients stochastiques (Ermoliev (1983)), et s'en sert pour évaluer la borne inférieure de la fonction de recours (i.e l'espérance des coûts de seconde étape).

Les méthodes ci-dessus sont toutes basées sur l'échantillonnage d'une distribution que l'on suppose connue. Cependant, il est souvent avantageux de construire des scénarios pour modéliser les aléas du problème plutôt que de les tirer au hasard.

Une méthode très simple pour discrétiser la distribution d'une variable aléatoire est de diviser l'espace en intervalles de probabilités égales et de leur assigner une valeur représentative (moyenne, médiane, etc.). Or, Miller et Rice (1983) démontrent que ce procédé aboutit à de mauvaises estimations des moments de la distribution. Ils suggèrent donc d'appliquer une méthode, appelée correspondance des moments et inspirée de la quadrature de Gauss, pour déterminer un ensemble de N paires de valeurs-probabilités qui respectent les $(2N-1)$ premiers moments. Cependant, les systèmes d'équations polynomiales engendrés par la quadrature de Gauss sont trop gros pour être résolus numériquement lorsque la dimension est supérieure à 1. Pour traiter le cas multidimensionnel, Hyland et Wallace (2001) proposent de générer des scénarios à l'aide de l'optimisation par la méthode des moindres carrés minimisant la distance entre les paramètres aléatoires désirés (moments, corrélations, événements extrémaux) et ceux découlant de l'arbre de scénarios. Hyland *et al.* (2003) démontrent qu'il est également possible d'implémenter une heuristique plus rapide basée sur la décomposition

de Cholesky de la matrice des corrélations et les transformations cubiques des distributions marginales pour obtenir une discrétisation possédant approximativement les mêmes quatre premiers moments et matrice de corrélations que la distribution originale.

La méthode de correspondance des moments n'est toutefois pas toujours appropriée. Keefer (1994) argumente qu'il existe plusieurs cas où une distribution est mal représentée par un ensemble fini de points qui possèdent exactement les mêmes premiers moments. Hochreiter et Pflug (2007) nous préviennent également des dangers de la méthode de correspondance des moments en présentant un exemple de quatre distributions possédant les mêmes quatre premiers moments, mais qui ont des comportements totalement opposés. En plus de mener à des solutions différentes au problème du vendeur de journaux (*newsvendor problem*), le moment d'ordre 5 d'une de ces distributions est infini tandis qu'il vaut 0 pour une autre. Finalement, Heyde (2010) démontre qu'il existe des distributions différentes en soi, mais possédant les mêmes moments pour tous les ordres.

On peut voir les scénarios comme étant les représentants d'un sous-espace du support de la distribution auxquels on assigne une probabilité. Il est donc naturel que les méthodes de partitionnement de données (*clustering analysis*) servent également à générer les scénarios. Gulpinar *et al.* (2004) développent une méthode hybride de génération de scénarios basée sur l'optimisation et le partitionnement de données. Le partitionnement de données est d'abord utilisé pour regrouper les noeuds de l'arbre de scénarios dont les poids sont ensuite déterminés par optimisation. L'avantage de l'algorithme est qu'il réduit considérablement le temps de convergence du programme d'optimisation puisque les noeuds ont déjà été choisis ; il ne reste à déterminer que les poids. Latorre *et al.* (2007) présentent quatre méthodes de génération de scénarios basées sur le partitionnement de données : le partitionnement de données conditionnel (*conditional clustering*), la méthode du gaz neuronal (*neural gas method*), le partitionnement de noeuds (*node clustering*) et le partitionnement progressif (*progressive clustering*). Parmi celles-ci, la méthode du gaz neuronal semble donner les meilleurs résultats malgré le fait qu'elle requiert l'ajustement d'un grand nombre de paramètres. La méthode choisie pour ce mémoire nous conduira également à utiliser des techniques de partitionnement de données.

L'erreur de quantification correspond à la différence entre les valeurs optimales de l'optimisation à partir de scénarios et à partir de la distribution réelle de la variable aléatoire. L'objectif est donc de minimiser cette erreur, que l'on souhaite le plus près de zéro possible. Pflug (2001) démontre qu'il existe une relation entre l'erreur de quantification et la distance de Wasserstein, et suggère que l'on devrait alors chercher à minimiser cette dernière par des

techniques de quantification optimale (aussi appelée discrétisation optimale). Il est possible de démontrer que la minimisation de la distance de Wasserstein est équivalente au problème de localisation d'entrepôts (*facility location problem*) et au problème des k-médianes en norme L_1 (voir Jain et Vazirani (2001)). Pflug (2001) propose un algorithme d'apprentissage compétitif basé sur la méthode des k-moyennes pour trouver les scénarios qui minimisent la distance de Wasserstein tandis que Hochreiter et Pflug (2007) utilisent plutôt un algorithme glouton de partitionnement de données. Lorsque la distribution est connue, Pages et Printems (2003) présentent un algorithme de quantification optimale par apprentissage compétitif (*Competitive Learning Vector Quantization* (CLVQ)). Ce dernier est valable pour toutes les normes de distance, mais demande l'ajustement de certains paramètres.

La génération de scénarios par quantification optimale nous pousse à chercher des méthodes efficaces de partitionnement de données. Likas *et al.* (2002) présente un algorithme glouton de partitionnement de données par sélection progressive des centres. L'algorithme de Chan *et al.* (2006) fonctionne de la manière inverse : un surplus de centres est d'abord choisi pour ensuite les éliminer un à un jusqu'à ce que l'on obtienne la quantité de partitions désirée. Chan *et al.* (2006) comparent leur algorithme à celui de Likas *et al.* (2002) et les résultats semblent indiquer que ce premier procure de meilleurs résultats en termes de mesure de distance et temps de convergence. Jain et Vazirani (2001) présente un algorithme qui permet de résoudre le problème des k-médianes avec un facteur d'approximation 3, c'est-à-dire avec une solution inférieure à 3 fois la solution optimale. Si l'on contraint les centres de la méthode des k-médianes à être positionnés sur les données, on obtient alors le problème des k-médoïdes. Ce dernier constitue une bonne approximation du premier et converge souvent plus rapidement (voir Velmurugan et Santhanam (2010)). Korupolu *et al.* (2000) proposent un algorithme qui résout le problème des k-médoïdes avec un facteur d'approximation (α, β), c'est-à-dire qui utilise βK centres pour trouver un partitionnement dont la solution est inférieure ou égale à α fois la solution optimale obtenue avec K centres. Indyk (1999) se sert de l'algorithme de Korupolu *et al.* (2000) sur un sous-ensemble des données pour résoudre le problème des k-médianes en un temps de convergence plus rapide ($O(n)$), mais avec un moins bon facteur d'approximation en contrepartie. Charikar *et al.* (2003) développent également une heuristique pour le problème des k-médianes avec un facteur d'approximation constant de 4, c'est-à-dire dont la solution est au plus 4 fois la solution optimale. Le temps de convergence des algorithmes de partitionnement peut être prohibitif lorsque le nombre de données à regrouper est trop grand. Chen (2009) propose donc un algorithme qui permet de réduire le nombre de données avant de procéder au partitionnement.

Il est parfois nécessaire d'utiliser des méthodes de réduction de scénarios lorsque l'arbre de scénarios est trop gros. On a déjà mentionné que ceci pouvait être fait par des méthodes de partitionnement de données (voir Gulpinar *et al.* (2004) ou Latorre *et al.* (2007)), mais il existe d'autres moyens. Carino *et al.* (1994) présentent une méthode de réduction de scénarios très simple qui permet de conserver l'espérance et la variance de la distribution. Kouwenberg (2001) teste la génération de scénarios par échantillonnage antithétique suivi de la méthode de réduction de Carino *et al.* (1994) et la correspondance des moments de Hyland et Wallace (2001) ; cette dernière donne des résultats légèrement meilleurs, mais les deux méthodes fonctionnent nettement mieux que l'échantillonnage pur dans les deux cas.

La génération de scénarios devient beaucoup plus difficile lorsque la dimension est élevée. Un moyen de remédier à ce problème est de générer les scénarios à l'aide des copules, qui permettent de traiter les distributions marginales et les corrélations séparément (Sutiene et Pranevicius (2007)). Lorsqu'on parle de corrélation entre deux variables aléatoires, il est souvent sous-entendu qu'il s'agit du coefficient de corrélation de Pearson, qui n'évalue que la dépendance linéaire entre celles-ci. Un avantage des copules est qu'elles tiennent également compte des dépendances non linéaires entre les variables. On doit d'abord générer les «scénarios de copules» en échantillonnant la distribution désirée, en utilisant une famille paramétrique de copules ou en les construisant directement à partir d'un programme d'optimisation (Kaut (2013)). Ceux-ci permettent ensuite de trouver les scénarios en associant les discrétisations de chaque distribution marginale. Kaut et Wallace (2011) ont testé l'utilisation des copules avec la méthode de correspondance des moments sur un problème de Valeur-à-Risque conditionnelle (*Conditional Value-at-Risk* ou *CVaR*). Ils obtiennent des solutions plus stables et moins biaisées lorsque les copules sont utilisées que lorsqu'elles ne le sont pas. La génération de scénarios par les copules sera testée dans le dernier chapitre.

L'optimisation stochastique est généralement plus avantageuse par rapport à l'optimisation déterministe lorsque la variance des variables aléatoires est élevée. Ainsi, pour les problèmes en dimension élevée, les distributions marginales les plus influentes sur le PS sont normalement celles possédant les plus grandes variances. Il peut alors être profitable de se servir de l'analyse par composantes principales (ACP), qui réduit la dimension en conservant une bonne partie de la variance et des covariances entre les variables aléatoires (voir Jolliffe (2002)). Le but de l'ACP n'est pas de trouver des scénarios possédant les mêmes variances marginales que la distribution originale, mais plutôt de considérer les dimensions possédant les plus grandes variances pour ensuite procéder à la génération de scénarios par une méthode quelconque. Jamshidian et Zhu (1996) utilisent l'ACP pour générer des scénarios en analyse

financière, tandis que Topaloglou *et al.* (2002) réduisent un problème stochastique de 15 à 7 variables par l'ACP tout en conservant 97 % de la variance. Ils suggèrent également de réajuster les composantes principales obtenues de manière à conserver l'espérance des variables.

On considère dans ce mémoire le problème d'optimisation stochastique de l'horaire du personnel dans un supermarché où la variable aléatoire correspond à la demande en employés. Legrain (2012) a établi une méthode de génération de scénarios pour ce problème qui divise la journée en 96 périodes en supposant que la demande en employés suit une loi log-normale pour chacune d'entre elles. Une fois les scénarios générés, on doit procéder à l'optimisation stochastique du problème (voir Birge et Louveaux (2011)). L'algorithme *L-shaped* est une méthode efficace qui génère des coupes d'optimalité et de réalisabilité (Benders (1962)) itérativement jusqu'à ce que la solution optimale soit atteinte. Pacqueau (2011) utilise une heuristique basée sur la méthode *L-shaped* afin de résoudre le problème de confection d'horaires pour les employés d'un supermarché qui donnent des résultats particulièrement favorables lorsque la variance de la variable aléatoire est élevée.

CHAPITRE 3

QUANTIFICATION OPTIMALE

3.1 Génération de scénarios

La première étape à suivre pour résoudre un problème d'optimisation stochastique est d'écrire le programme stochastique (PS), composé de la fonction objectif, des contraintes, des variables de décision et des variables aléatoires spécifiques à notre problème. Le PS se distingue du programme déterministe (PD) par la présence de variables aléatoires qui complexifie le problème et augmente le temps de résolution numérique. Le PD ne contient pas de variables aléatoires ou les approxime par une valeur unique, qui correspond généralement à la moyenne ou la médiane. La grande majorité des PS ne peuvent être résolus lorsque le support des variables est infini, qu'elles soient discrètes ou continues. Il est donc nécessaire de procéder à la génération de scénarios pour quantifier la distribution du vecteur aléatoire, c'est-à-dire le représenter par un ensemble fini de réalisations nommées scénarios ayant chacun une probabilité qui lui est associé. Le programme stochastique peut ensuite être résolu par un algorithme d'optimisation tel que la méthode *L-shaped*. À moins que la méthode de génération soit défectueuse, plus il y a de scénarios, meilleure sera la représentation des variables aléatoires. En revanche, même s'il est spécifique à chaque problème, le temps de calcul numérique augmente toujours avec la quantité de scénarios. Il faut donc juger du nombre de scénarios que l'on désire générer pour chaque problème en fonction du temps de calcul et de la précision recherchés.

Les programmes stochastiques dépendent parfois de la réalisation de plusieurs vecteurs aléatoires successifs. La décision de première étape est donc prise avant toute observation d'un vecteur aléatoire tandis que les autres doivent être prises séquentiellement après chaque réalisation d'un vecteur aléatoire. On parle alors de programmation stochastique à plusieurs étapes dont les scénarios sont souvent représentés pas un arbre comme celui de la figure 3.1. Les arcs correspondent à des réalisations du vecteur aléatoire tandis que les noeuds représentent les décisions du problème. Un scénario correspond donc à un chemin allant du noeud initial jusqu'à un noeud terminal. Il s'agit ici d'un arbre à 3 étapes de décision t_1, t_2 et t_3 comportant 5 scénarios $\mathbf{v}_1, ..., \mathbf{v}_5$. Les probabilités associées à chaque scénario correspondent au produit des probabilités conditionnelles. Supposons par exemple que le premier vecteur aléatoire prenne des valeurs dans l'ensemble $\{\mathbf{v}(i), 1 \leq i \leq 2\}$ et que les réalisa-

tions du 2^e proviennent de l'ensemble $\{\mathbf{v}(i,j) : 1 \leq i \leq 2, 1 \leq j \leq 3\}$. Si $p_1 = P[\mathbf{v}(1)]$ et $p_2 = P[\mathbf{v}(1,2)|\mathbf{v}(1)]$, alors la probabilité associée au scénario $\mathbf{v}_2 = (\mathbf{v}(1), \mathbf{v}(1,2))$ est $p_1 p_2$.

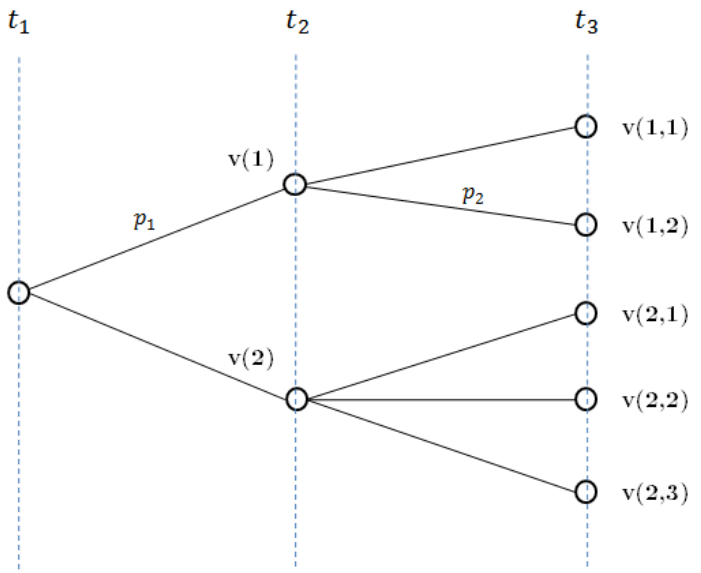

Figure 3.1 Arbre de scénarios à 3 étapes de décision.

Dans ce mémoire, on ne traite que des PS à deux étapes. Une décision de 1^{re} étape doit donc être prise avant la réalisation du vecteur aléatoire et la décision de 2^e étape, appelée recours, est prise après. Le tableau 3.1 présente la notation qui sera utilisée au cours de ce mémoire.

Tableau 3.1 Notation utilisée.

Notation	Signification
D	Dimension du problème
$\mathbf{U}$	Vecteur aléatoire à D dimensions
U_i	Variable aléatoire de la i^e dimension de $\mathbf{U}$, $1 \leq i \leq D$
K	Nombre de scénarios générés
$\mathbf{V}$	Distribution des scénarios
$\mathbf{v}_j$	j^e scénario, $1 \leq j \leq K$
p_j	Probabilité de réalisation du scénario $\mathbf{v}_j$, $1 \leq j \leq K$

On considère dans ce mémoire les problèmes de minimisation, mais l'étude peut facilement se généraliser aux problèmes de maximisation. Avec la notation définie dans le tableau ci-

dessus, le PS peut s'écrire

$$\min_{x \in X} \; Z(x) = \int z(x, \mathbf{u}) \; f_{\mathbf{U}}(\mathbf{u}) \; d\mathbf{u} \tag{3.1}$$

où $f_{\mathbf{U}}(\mathbf{u})$ est la fonction de densité de $\mathbf{U}$, x représente les variables de décision [1], X est le domaine réalisable du PS, $z(x, \mathbf{u})$ est la fonction de coût pour une réalisation $\mathbf{u}$ du vecteur aléatoire donnée et Z est la fonction objectif. La fonction objectif correspond donc à l'espérance de la fonction de coût, qui dépend à la fois des variables de décision et de l'événement aléatoire. On notera une solution optimale x^* et sa valeur optimale Z^*.

La formulation classique du PS à deux étapes s'obtient en réécrivant la fonction de coût comme

$$z(x, \mathbf{u}) = g(x) + Q(x, \mathbf{u}) \; , \tag{3.2}$$

où g est une fonction qui dépend de la décision de 1$^{\text{re}}$ étape et $Q(x, \mathbf{u})$ est la valeur optimale du problème de seconde étape

$$Q(x, \mathbf{u}) = \min_{y} \; q(y, \mathbf{u}) \tag{3.3}$$

$$\text{s.c.} \quad \mathrm{T}(\mathbf{u}) \; x - \mathrm{W}(\mathbf{u}) \; y = h(\mathbf{u}) \tag{3.4}$$

$$y \geq 0$$

où q une fonction qui dépend de la décision de seconde étape et de la réalisation $\mathbf{u}$. Les matrices $\mathrm{T}(\mathbf{u})$ et $\mathrm{W}(\mathbf{u})$ ainsi que le vecteur $h(\mathbf{u})$ peuvent tous dépendre de $\mathbf{u}$. On cherche alors à résoudre

$$\min_{x \in X} \; Z(x) = g(x) + E_{\mathbf{U}} \left[Q(x, \mathbf{U}) \right], \tag{3.5}$$

où $E_{\mathbf{U}}$ est l'espérance sur la distribution de $\mathbf{U}$. Cependant, ce PS ne peut généralement pas être résolu puisque $\mathbf{U}$ peut prendre une infinité de valeurs. La variable aléatoire $\mathbf{U}$ doit donc être remplacée par un nombre fini de scénarios et (3.1) devient

$$\min_{x \in X} \; \tilde{Z}(x; \mathbf{V}) = \sum_{j=1}^{K} \; z(x, \mathbf{v}_j) \; p_j \tag{3.6}$$

Une valeur optimale ainsi modifiée de (3.6) sera notée $\tilde{x}^*$. La variable aléatoire $\mathbf{U}$ correspond à la distribution réelle tandis que $\mathbf{V}$ est une approximation de la réalité. En théorie, il

1. Pour plus de clarté, on notera x les variables de décisions, même si elles correspondent à un vecteur.

est possible d'obtenir une approximation de $\mathbf{U}$ aussi fine que désiré à l'aide d'un très grand nombre de points discrets avec des probabilités qui leur sont associées. Cependant, plus le nombre de scénarios est grand, plus le temps de convergence numérique le sera également. Ainsi, pour un temps de résolution numérique maximal $\bar{t}$ spécifié, on ne pourra considérer qu'un ensemble restreint de réalisations dépendant de la complexité du PS. Le coût *réel* obtenu en optimisant le PS avec les scénarios est

$$Z(\tilde{x}^*) = \int z(\tilde{x}^*, \mathbf{u})\ f_{\mathbf{U}}(\mathbf{u})\ d\mathbf{u} \tag{3.7}$$

et l'erreur de cette approximation est donnée par

$$e(Z, \tilde{Z}) = Z(\tilde{x}^*) - Z(x^*). \tag{3.8}$$

Il n'est évidemment pas possible de mesurer cette erreur puisque cela impliquerait que l'on soit capable de résoudre (3.1) et trouver $Z(x^*)$, auquel cas on n'aurait pas besoin de générer des scénarios. Notons que $e(Z, \tilde{Z}) \geq 0$, puisque $Z(\tilde{x}^*) \geq Z(x^*)$.

Le but de la génération de scénarios est donc de trouver une distribution $\mathbf{V}$ permettant de résoudre (3.6) en un temps raisonnable et qui minimise $e(Z, \tilde{Z})$. Pour un PS donné, le temps de convergence numérique dépend du nombre de scénarios. On pourrait donc reformuler l'objectif comme étant de trouver une distribution $\mathbf{V}$ possédant un nombre restreint de scénarios et qui minimise $e(Z, \tilde{Z})$.

3.2 Justification de la méthode

Comme il a déjà été dit, il existe plusieurs méthodes de génération de scénarios. En supposant que l'on connaît la distribution de $\mathbf{U}$, une des méthodes les plus simples consiste à échantillonner K valeurs de cette distribution et de les utiliser directement comme scénarios. Leurs probabilités respectives sont ensuite calculées et correspondent aux proportions des échantillons (sur un grand nombre de tirages) dont la distance est la plus près de chaque scénario. Cette procédure sera appelée *échantillonnage pur* et elle nous servira d'étalon de comparaison avec les autres méthodes. Nous désirons développer une technique plus sophistiquée qui permettra d'obtenir une erreur $e(Z, \tilde{Z})$ considérablement inférieure à celle de l'échantillonnage pur.

Il serait également possible d'utiliser des méthodes Monte Carlo telles que l'échantillon-

nage préférentiel pour réduire la variance de l'estimateur. Cependant, les méthodes d'échantillonnage sont plus efficaces lorsque plusieurs échantillons sont tirés afin d'obtenir une faible variance de l'estimateur. Or, certains programmes stochastiques plus complexes ne peuvent considérer qu'un ensemble restreint de scénarios (25, par exemple) pour conserver des temps de convergence raisonnable. De plus, nous verrons que les méthodes utilisées sont développées dans le but spécifique de réduire la borne supérieure sur l'erreur d'approximation et donnent de meilleurs résultats que l'échantillonnage pur. Il est donc fort probable qu'elles offrent également de meilleurs résultats que les autres techniques d'échantillonnage, même si elles permettent de réduire la variance de l'estimateur. Par conséquent, nous préférerons développer des algorithmes qui construisent les scénarios itérativement plutôt que de les échantillonner directement à partir de la distribution.

Une technique de génération de scénarios très utilisée est la méthode de *correspondance des moments* (ou méthode des moments) de Kaut et Wallace (2011). En supposant que l'on connaît la distribution de **U**, le principe de cette méthode est de trouver une nouvelle distribution finie **V** qui possède les mêmes premiers moments et co-moments que **U**. Cependant, certaines distributions possédant les mêmes premiers moments et co-moments peuvent être très différentes les unes par rapport aux autres. Hochreiter et Pflug (2007) donnent en exemple les quatre distributions suivantes :

1. Uniforme sur l'intervalle [-2.44949, 2.44949]
2. Distribution mixte de gaussiennes $\mathcal{N}(1.244666, 0.450806)$ et $\mathcal{N}(-1.244666, 0.450806)$ avec probabilités 1/2 chacune
3. Une distribution discrète D_1 avec les valeurs et probabilités suivantes :

Valeur	-2.0395	-0.91557	0	0.91557	2.0395
Probabilité	0,2	0,2	0,2	0,2	0,2

4. Une distribution discrète D_2 avec les valeurs et probabilités suivantes :

Valeur	-3.5	-1.4	0	1.4	3.5
Probabilité	0,013	0,429	0,116	0,429	0,013

Les 4 premiers moments de ces distributions ont des valeurs identiques respectivement égales à 0, 2 , 0 et 7,2. On constate cependant à partir de la figure 3.2 que les fonctions de masse ou de densité sont très différentes. Le problème du vendeur de journaux a été résolu par Hochreiter et Pflug (2007) en utilisant tour à tour chacune des distributions pour représenter la variable aléatoire. Ce dernier consiste à déterminer la quantité optimale de journaux

qu'un vendeur devrait acheter afin de maximiser ses profits, étant donné des prix d'achat, de vente, de rachat fixés ainsi que la distribution aléatoire de la demande des clients. Comme on pouvait s'y attendre, les valeurs optimales obtenues diffèrent largement d'une distribution à l'autre.

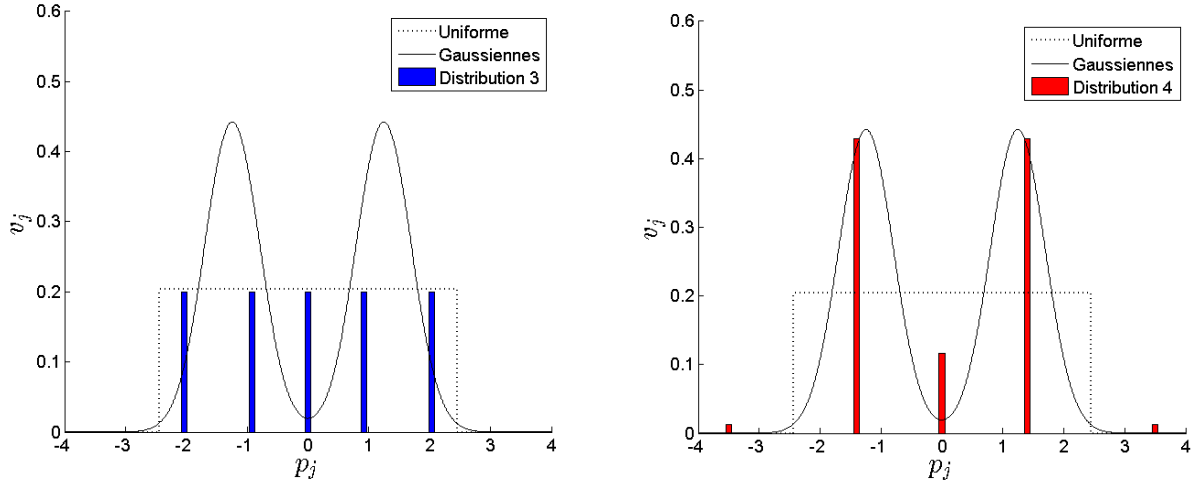

Figure 3.2 Fonctions de masse de distributions possédant 4 premiers moments identiques

Il existe plusieurs autres exemples de distributions ayant certains moments en commun, mais dont le comportement est très différent. Par exemple, Heyde (2010) démontre même que deux distributions intrinsèquement différentes peuvent posséder les mêmes moments pour tous les ordres. Il est donc clair que la méthode des moments peut mener à des résultats désastreux. Par conséquent, il n'est pas souhaitable de concentrer nos efforts afin de faire coïncider les moments des scénarios à ceux de la distribution originale et la méthode de correspondance des moments ne sera pas considérée dans ce mémoire.

Nous utiliserons plutôt une méthode de génération de scénarios inspirée de Pflug (2001) appelée *quantification optimale*. Nous avons vu qu'il était impossible d'obtenir une expression pour $e(Z, \tilde{Z}) = Z(\tilde{x}^*) - Z(x^*)$, puisqu'on ne connaît pas de solution optimale réelle x^*. Notons que si on était en mesure d'obtenir x^*, la génération de scénarios ne serait pas nécessaire en premier lieu. La méthode de quantification optimale cherche donc à minimiser une borne supérieure sur $e(Z, \tilde{Z})$, plutôt que l'erreur elle-même. On fait donc l'hypothèse selon laquelle la minimisation de la borne supérieure engendre des valeurs pour $e(Z, \tilde{Z})$ assez près du minimum. Nous verrons que cette méthode nous conduira également vers l'utilisation de

techniques de partitionnement de données.

3.3 Présentation de la méthode

La méthode de quantification optimale fait appel à plusieurs notions de topologie. Ainsi, nous commençons par présenter quelques résultats sur les métriques qui nous serviront à trouver une borne supérieure sur $e(Z, \tilde{Z})$.

Définition 3.1. *Une* **métrique** *est une application* $\rho : \mathbb{E} \times \mathbb{E} \to [0, \infty)$ *qui satisfait les trois propriétés suivantes :*

1. *identité :* $\rho(x, y) = 0 \iff x = y$
2. *symétrie :* $\rho(x, y) = \rho(y, x)$
3. *inégalité du triangle :* $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$,

$\forall\ x,\ y$ *et* $z \in \mathbb{E}$. *Une métrique* ρ *associée à un ensemble* $\mathbb{E}$ *forme un* **espace métrique** $(\mathbb{E}, \rho)$.

L'espace métrique le mieux connu est sans doute l'espace euclidien $(\mathbb{R}^{\mathrm{D}}, \|\cdot\|_2)$, où $\|\cdot\|_2$ correspond à la distance euclidienne

$$\|\mathbf{u}\|_2 = \left(\sum_{i=1}^{D} |u_i|^2 \right)^{1/2}$$

et $\mathbf{u} \in \mathbb{R}^{\mathrm{D}}$. La distance euclidienne est un cas particulier ($p = 2$) de la distance induite par la norme L_p

$$\|\mathbf{u}\|_p = \left(\sum_{i=1}^{D} |u_i|^p \right)^{\min\{1, 1/p\}}, \tag{3.9}$$

où p est un nombre réel strictement positif. Dans le cadre de cette recherche, on utilisera plutôt les métriques *probabilistes* (Rachev *et al.* (2013)). La définition d'une métrique probabiliste est un peu plus abstraite et fait appel à certaines notions de topologie qui ne seront pas traitées dans le cadre de ce mémoire. Sans entrer dans les détails, on peut définir une métrique probabiliste comme une mesure de distance entre les fonctions de répartition de deux distributions. Par exemple, la mesure suivante correspond à une métrique entre la valeur absolue des moments d'ordre $p \geq 1$ des distributions U et V :

$$\mathbf{MOM}_p(\mathrm{U},\mathrm{V}) = \left| \ (E|\mathrm{U}^p|)^{1/p} - (E|\mathrm{V}^p|)^{1/p} \ \right| . \tag{3.10}$$

Nous utiliserons plutôt une métrique probabiliste nommée la distance de Wasserstein (d'après

Wasserstein (1969)), qui nous permettra de trouver une borne supérieure sur l'erreur de quantification.

Définition 3.2. *Soit* $\rho : \mathbb{E} \times \mathbb{E} \to [0, \infty)$ *une métrique et* $\mathbf{U}$, $\mathbf{V}$ *des distributions avec fonctions de répartition* $F_{\mathbf{U}}$ *et* $F_{\mathbf{V}}$ *respectivement. La* **distance de Wasserstein** *est définie par*

$$d_1(\mathbf{U},\mathbf{V}) = \inf_{F_{\mathbf{U},\mathbf{V}}} E[\rho(\mathbf{U},\mathbf{V})] \ . \tag{3.11}$$

On remarque que l'infimum dans (3.11) est pris sur l'ensemble des fonctions de répartition conjointes $F_{\mathbf{U},\mathbf{V}}$ ayant des fonctions de répartition «marginales» $F_{\mathbf{U}}$ et $F_{\mathbf{V}}$ fixées. On pourrait démontrer que la distance de Wasserstein est effectivement une métrique probabiliste. Nous présentons maintenant un théorème fondamental qui donne une formulation duale de la distance de Wasserstein. Le théorème a d'abord été démontré par Kantorovich et Rubinstein (1958) dans le cas où $(\mathbb{E}, \rho)$ forme un espace métrique compact. Il a ensuite été généralisé par Dudley (1976) et Levin et Milyutin (1979) en relaxant entre autres la condition que ρ soit une métrique.

Théorème 3.1. (Théorème de Kantorovich et Rubinstein (1958)) *Soit* $(\mathbb{E}, \rho)$ *un espace métrique séparable*[2] *et* $\mathbf{U}$ *et* $\mathbf{V}$ *les distributions associées à la distance de Wasserstein. Soit* ψ *une fonction telle que*

$$|\psi(\mathbf{u}) - \psi(\mathbf{v})| \leq \rho(\mathbf{u},\mathbf{v}) \ , \tag{3.12}$$

alors

$$d_1(\mathbf{U},\mathbf{V}) = \sup_{\psi} \left| \int \psi(\mathbf{u}) \ dF_{\mathbf{U}}(\mathbf{u}) - \int \psi(\mathbf{v}) \ dF_{\mathbf{V}}(\mathbf{v}) \right| \ . \tag{3.13}$$

Preuve On réfère le lecteur intéressé à consulter Huber (2005) ou Rachev *et al.* (2013) pour des démonstrations récentes.

Le théorème de Kantorovich-Rubinstein nous servira maintenant à trouver une borne supérieure sur l'erreur de discrétisation $e(Z, \tilde{Z})$ définie par (3.8). Nous commençons par considérer les variables aléatoires en une seule dimension et généraliserons par la suite au cas multidimensionnel.

3.3.1 Cas unidimensionnel

Nous suivons le raisonnement présenté dans Pflug (2001) pour déterminer une borne supérieure sur $e(Z, \tilde{Z})$. On énonce d'abord le lemme suivant tiré de Pflug (2001) et dont la démonstration est présentée dans son ouvrage.

2. On supposera que cette condition est toujours respectée. En fait, on se limitera au cas où $\mathbb{E} = \mathbb{R}^D$.

Lemme 3.1.

$$e(Z, \tilde{Z}) \leq 2 \sup_{x \in X} \left| Z(x) - \tilde{Z}(x) \right| \tag{3.14}$$

Pflug (2001) présente également le résultat suivant, que nous énonçons sous forme de proposition. La démonstration est assez triviale et a donc été omise de son article, mais nous ajoutons tout de même notre version de la preuve par souci de clarté. On notera u (resp. v) une réalisation de la variable aléatoire U (resp. V).

Proposition 3.1. *Supposons que la fonction de coût $z(x, \cdot)$ soit uniformément Lipschitz d'ordre 1 et de constante $\overline{L}_1$, c'est-à-dire*

$$\inf \{L : |z(x,u) - z(x,v)| \leq L \cdot |u - v| \ \forall \ u \in \Xi_U, v \in \Xi_V\} \leq \overline{L}_1 \quad \forall \ x \in X \ , \tag{3.15}$$

où Ξ_U et Ξ_V correspondent aux supports des variables aléatories U *et* V *respectivement et X représente le domaine réalisable du PS. Si* U *est approximée par une distribution* V*, alors*

$$e(Z, \tilde{Z}) \leq 2 \ \overline{L}_1 \cdot d_1(\mathrm{U},\mathrm{V}) \tag{3.16}$$

Preuve Posons $\psi(x,u) = z(x,u)/\overline{L}_1$. Puisque $z(x, \cdot)$ est uniformément Lipschitz de constante $\overline{L}_1$ par hypothèse, $\psi(x, \cdot)$ est uniformément Lipschitz de constante 1

$$|\psi(x,u) - \psi(x,v)| \leq |u - v| \quad \forall \ x \in X \tag{3.17}$$

et

$$\begin{aligned} \frac{1}{\overline{L}_1} \sup_x \left| Z(x) - \tilde{Z}(x) \right| &= \sup_x \left| \int \frac{z(x,u)}{\overline{L}_1} f_U(u)du - \int \frac{z(x,v)}{\overline{L}_1} f_V(v)dv \right| \\ &= \sup_x \left| \int \psi(x,u) f_U(u)du - \int \psi(x,v) f_V(v)dv \right| \\ &\leq \sup_{\psi(x,\cdot)} \left| \int \psi(x,u) f_U(u)du - \int \psi(x,v) f_V(v)dv \right| \ . \end{aligned} \tag{3.18}$$

Par le théorème de Kantorovich-Rubinstein, (3.18) est égale à d_1(U,V) avec $\rho(u,v) = |u - v|$. On obtient donc

$$\frac{1}{\overline{L}_1} \sup_x \left| Z(x) - \tilde{Z}(x) \right| \leq d_1(\mathrm{U},\mathrm{V})$$

et l'inégalité (3.16) découle du lemme 3.1. □

Nous avons donc atteint l'objectif de déterminer une borne supérieure sur l'erreur, donnée par (3.16). La constante uniforme de Lipschitz $\overline{L}_1$ est entièrement déterminée par la fonction

de coût $z(x,\cdot)$ et fait le lien entre cette borne et le programme stochastique. Il n'est donc pas possible de réduire la valeur de $\overline{L}_1$ à moins de modifier le PS. La minimisation de la borne supérieure revient donc à minimiser d_1(U,V). D'un autre côté, le PS n'a aucune influence sur la distance de Wasserstein, qui ne dépend que des fonctions de répartition marginale de U et V. La méthode de quantification optimale revient donc à chercher une distribution discrète V qui minimise d_1(U,V), où U représente la distribution à quantifier et les réalisations $v_j,\ j = 1, ..., K$, correspondent aux scénarios. La minimisation de d_1(U,V) n'est pas toujours évidente, mais la proposition suivante tirée de Pflug (2001) nous guidera vers une méthode pour y parvenir.

Proposition 3.2. *Soit* U *une distribution avec fonction de répartition* F_{U} *et* V *une distribution discrète de support fini qui prend les valeurs* $v_j,\ j = 1, ..., K$, *avec probabilité* p_j. *On suppose que les réalisations* v_j *sont classées en ordre croissant* $v_1 < ... < v_K$. *Soit* $\rho(u,v) = |u - v|$ *la métrique associée à la distance de Wasserstein. Les probabilités* p_j *qui minimisent* d_1(U,V) *sont*

$$p_j = \int_{\frac{v_{j-1}+v_j}{2}}^{\frac{v_j+v_{j+1}}{2}} f_{\mathrm{U}}(u)du \ . \tag{3.19}$$

De plus, la distance de Wasserstein est donnée par

$$d_1(\mathrm{U},\mathrm{V}) = \sum_{j=1}^{K} \int_{\frac{v_{j-1}+v_j}{2}}^{\frac{v_j+v_{j+1}}{2}} |u - v_j| \ f_{\mathrm{U}}(u) \ du \ , \tag{3.20}$$

où $v_0 = -\infty$ *et* $v_{K+1} = \infty$. *Le supremum de* (3.13) *est atteint pour* $\psi(u) = \min_j |u - v_j|$.

Preuve Selon Pflug (2001), la démonstration découle aisément du théorème présenté dans Vallender (1974). Nous donnons ici une autre version de la preuve, qui ne fait pas référence au théorème de Vallender (1974).

On utilise directement la définition de la distance de Wasserstein

$$\begin{aligned} d_1(\mathrm{U},\mathrm{V}) &= \min_{F_{\mathrm{U},\mathrm{V}}} E[\rho(\mathrm{U},\mathrm{V})] \\ &= \min_{F_{\mathrm{U},\mathrm{V}}} E\left[|\mathrm{U} - \mathrm{V}|\right] \\ &= \min_{F_{\mathrm{U},\mathrm{V}}} \sum_{j=1}^{K} \int |u - v_j| \ f_{\mathrm{U}|\mathrm{V}}(u|v_j) \ du \ p_j \ . \end{aligned} \tag{3.21}$$

Il est clair que d_1(U,V) sera minimal si on peut choisir une distribution conjointe qui associe

chaque réalisation de U à la valeur v_j la plus près avec probabilité 1, i.e.

$$P\left[\frac{v_{j-1}+v_j}{2} \leq \mathrm{U} \leq \frac{v_j+v_{j+1}}{2} \mid \mathrm{V}=v_j\right]=1 \tag{3.22}$$

$$\text{et} \qquad P\left[\mathrm{V}=v_j \mid \frac{v_{j-1}+v_j}{2} \leq \mathrm{U} \leq \frac{v_j+v_{j+1}}{2}\right]=1\ , \tag{3.23}$$

où $v_0=-\infty$ et $v_{K+1}=\infty$. Par le théorème de Bayes, on obtient

$$\begin{aligned}
\frac{1}{p_j}\cdot P\left[\mathrm{V}=v_j \mid \frac{v_{j-1}+v_j}{2} \leq \mathrm{U} \leq \frac{v_j+vj+1}{2}\right]\ P\left[\frac{v_{j-1}+v_j}{2} \leq \mathrm{U} \leq \frac{v_j+vj+1}{2}\right]&=1\\
\iff \quad p_j=P\left[\frac{v_{j-1}+v_j}{2} \leq \mathrm{U} \leq \frac{v_j+vj+1}{2}\right]&\\
p_j=\int_{\frac{v_{j-1}+v_j}{2}}^{\frac{v_j+v_{j+1}}{2}} f_{\mathrm{U}}(u)du\ .&
\end{aligned} \tag{3.19}$$

L'équation (3.21) devient alors

$$\begin{aligned}
d_1(\mathrm{U},\mathrm{V}) &= \sum_{j=1}^{K}\int_{\frac{v_{j-1}+v_j}{2}}^{\frac{v_j+v_{j+1}}{2}} |u-v_j|\ f_{\mathrm{U}|\mathrm{V}}(u|v_j)\ du\ p_j\\
&= \sum_{j=1}^{K}\int_{\frac{v_{j-1}+v_j}{2}}^{\frac{v_j+v_{j+1}}{2}} |u-v_j|\ f_{\mathrm{U},\mathrm{V}}(u,v_j)\ du\\
&= \sum_{j=1}^{K}\int_{\frac{v_{j-1}+v_j}{2}}^{\frac{v_j+v_{j+1}}{2}} |u-v_j|\ f(u)\ du\ .
\end{aligned} \tag{3.20}$$

On peut facilement démontrer que l'on obtient l'équation (3.20) en remplaçant $\psi(u)=\min_j |u-v_j|$ dans (3.13). Par le théorème de Kantorovich-Rubinstein, $\psi(u)$ est donc la fonction qui maximise la forme duale de la distance de Wasserstein, définie par (3.13). □

L'équation (3.19) nous permet de trouver les probabilités p_j associées à chaque scénario v_j qui minimisent la distance de Wasserstein. Cependant, la recherche des valeurs optimales v_j^* pour les scénarios est moins évidente. En effet, on ne peut pas espérer les trouver en dérivant directement (3.20) par rapport à v_j, à cause de leur présence dans les intégrales. On sera donc contraint de développer un algorithme qui converge vers v_j^* basé sur la proposition suivante.

Proposition 3.3. *Soit* U *une distribution avec fonction de répartition* F_{U}*. Supposons que* $\mathbb{R}$ *est divisé en intervalles* (α_j,α_{j+1})*,* $j=1,...,K$*, où les* α_j *sont des constantes telles que* $-\infty=\alpha_0<\alpha_1<...<\alpha_{K-1}<\alpha_K=\infty$*. On cherche à associer une valeur* v_j *à chaque*

intervalle (α_j, α_{j+1}) *de manière à minimiser*

$$g(v_0, ..., v_K) = \sum_{j=1}^{K} \int_{\alpha_j - 1}^{\alpha_j} |u - v_j| \ f_{\mathrm{U}}(u) \ du \ . \tag{3.24}$$

Les valeurs v_j^* *qui minimisent la fonction* g *correspondent aux médianes des intervalles* (α_j, α_{j+1}), $j = 1, ...K - 1$.

Preuve Considérons un indice $j = j_0$. En dérivant g par rapport à v_{j_0}, on obtient

$$\begin{aligned}
0 &= \frac{\partial g(v_0, ..., v_K)}{\partial v_{j_0}} \\
&= \frac{\partial}{\partial v_{j_0}} \sum_{j=1}^{K} \int_{\alpha_{j-1}}^{\alpha_j} |u - v_j| \ f_{\mathrm{U}}(u) \ du \\
&= \frac{\partial}{\partial v_{j_0}} \int_{\alpha_{j_0 - 1}}^{\alpha_{j_0}} |u - v_{j_0}| \ f_{\mathrm{U}}(u) \ du \\
&= \int_{\alpha_{j_0 - 1}}^{\alpha_{j_0}} \frac{\partial}{\partial v_{j_0}} |u - v_{j_0}| \ f_{\mathrm{U}}(u) \ du \\
&= \int_{\alpha_{j_0 - 1}}^{\alpha_{j_0}} sgn(u - v_{j_0}) \ f_{\mathrm{U}}(u) \ du \\
&= \int_{\alpha_{j_0 - 1}}^{v_{j_0}} (-1) \ f_{\mathrm{U}}(u) \ du + \int_{\alpha_{v_{j_0}}}^{\alpha_{j_0}} (1) \ f_{\mathrm{U}}(u) \ du \\
&\implies P[\alpha_{j_0 - 1} < U < v_{j_0}] = P[v_{j_0} < U < \alpha_{j_0}] \ .
\end{aligned} \tag{3.25}$$

L'équation ci-dessus définit la médiane de l'intervalle $(\alpha_{j_0}, \alpha_{j_0+1})$. On arriverait évidemment au même résultat pour tout j, donc $\nabla g(v_0^*, ..., v_K^*) = \overrightarrow{0}_K$, où v_j^*, $j = 1, ..., K$ correspondent aux médianes de chaque intervalle. □

L'analyse de notre problème de génération de scénarios nous a conduits à choisir la métrique $\rho(u, v) = |u - v|$. On pourrait démontrer de manière analogue à la preuve ci-dessus que si l'on choisit plutôt la distance $\rho(u, v) = |u - v|^2$, alors les scénarios optimaux correspondent aux moyennes de chaque intervalle (α_j, α_{j+1}). Ce résultat est à la source d'un algorithme développé par Lloyd (1982) initialement conçu pour traiter le problème des *k-moyennes*, où l'on cherche à discrétiser une distribution continue avec la métrique $\rho(u, v) = |u - v|^2$. Lorsque la fonction de densité est lisse et positive, alors il a été démontré que l'algorithme de Lloyd converge effectivement vers le minimum global du problème des k-moyennes (Du *et al.* (2006)). Nous en donnons ici la version modifiée pour le problème des *k-médianes*, associé à la métrique $\rho(u, v) = |u - v|$.

Algorithme de Lloyd (continu)

1. (*Initialisation*) On choisit un paramètre d'arrêt ϵ arbitrairement petit. On échantillonne K valeurs à partir de la distribution U et on les classe en ordre croissant $u_1 < ... < u_K$. On pose $v_j^0 = u_j,\ j = 1, ..., K$, et $s = 0$.
2. (*Intervalles*) On trouve les intervalles $\left(\frac{v_{j-1}^s + v_j^s}{2}, \frac{v_j^s + v_{j+1}^s}{2}\right)$, où $v_0 = -\infty$ et $v_{K+1} = \infty$.
3. (*Médianes*) On calcule les médianes $\nu_j^{1/2}$ de chaque intervalle trouvé à l'étape 2 et on pose $v_j^{s+1} = \nu_j^{1/2}$.
4. (*Critère d'arrêt*) Si $|v_j^{s+1} - v_j^s| < \epsilon\ \forall\ j$, on choisit $v_j^* = v_j^{s+1}$. Sinon, on incrémente s de 1 unité ($s \leftarrow s + 1$) et on retourne à la 2^e étape [3].

Pour évaluer la borne sur l'erreur (3.16), on doit calculer la distance de Wasserstein et la constante uniforme de Lipschitz. Lorsque la distribution U est continue et connue, l'algorithme de Lloyd représente un bon choix pour minimiser d_1(U,V) puisqu'il converge vers le minimum global. La probabilité de chaque scénario v_j correspond évidemment à $P\left[\frac{v_{j-1}+v_j}{2} < \mathrm{U} < \frac{v_j+v_{j+1}}{2}\right]$. Cette méthode n'est toutefois pas aussi efficace dans le cas multidimensionnel. Non seulement elle ne garantit plus l'atteinte du minimum global, mais elle devient plus lourde numériquement puisque le calcul des médianes correspond alors à des intégrations multiples sur des sous-espaces non triviaux de $\mathbb{R}^D$. On pourrait néanmoins démontrer qu'il génère une suite de solutions non croissantes (i.e. $d_1(\mathbf{U}, \mathbf{V}^{s+1}) \leq d_1(\mathbf{U}, \mathbf{V}^s)$) et converge donc vers un minimum local. On se concentrera sur le développement de méthodes alternatives pour minimiser la distance de Wasserstein au prochain chapitre.

Pour le calcul de $\overline{L}_1$, deux difficultés peuvent subvenir. Premièrement, il n'est pas toujours possible de trouver la constante uniforme de Lipschitz, surtout lorsque le PS est complexe. Cependant, la valeur de $\overline{L}_1$ est seulement utile pour mesurer la borne supérieure sur l'erreur. Si l'on n'est pas en mesure de la calculer, on peut tout de même procéder à la génération de scénarios en minimisant la distance de Wasserstein. Deuxièmement, nous avons supposé que la fonction de coût est uniformément Lipschitz d'ordre 1, alors que ce n'est pas toujours le cas en réalité. Si elle n'est pas Lipschitz du tout, alors la justification théorique de la méthode de quantification optimale ne tient plus. Par contre, si $z(x, \cdot)$ est uniformément Lipshitz d'ordre $p \in \mathbb{N}$, $p > 1$, i.e.

$$L_p(z(x,\cdot)) = \inf\left\{L : |z(x,u) - z(x,v)| \leq L \cdot |u - v|^p\right\} \leq \overline{L}_p \quad \forall\ x \in X\ , \tag{3.26}$$

3. Il est aussi possible d'arrêter l'algorithme après un nombre fixe d'itérations. C'est ce que nous avons fait dans le dernier chapitre.

alors il n'est pas possible d'utiliser directement la distance de Wasserstein pour représenter la borne supérieure sur l'erreur pour notre problème. En effet, la condition (3.12) du théorème de Kantorovich-Rubinstein tel que présenté ne serait plus respectée dans ce cas. Si l'on pose $\psi(x,u) = z(x,u)/\overline{L}_p$, on obtient

$$\psi(x,u) - \psi(x,v) \leq |u-v|^p \ , \tag{3.27}$$

mais la fonction $\rho(u,v) = |u-v|^p$ avec $p > 1$ n'est pas une métrique puisqu'elle ne respecte plus l'inégalité du triangle. Pflug (2001) démontre toutefois qu'en utilisant une transformation χ telle que

$$\chi_p(u) = |u|^p \ sgn(u) \tag{3.28}$$

$$\chi_{1/p}(u) = |u|^{1/p} \ sgn(u) \ , \tag{3.29}$$

alors le problème peut être réduit à celui de minimiser la distance de Wasserstein. Hochreiter et Pflug (2007) expliquent clairement les étapes pour minimiser la distance de Wasserstein en appliquant la transformation ci-dessus.

Notons que même si la fonction de coût est Lipschitz d'ordre $p > 1$, il est toujours possible de générer les scénarios en minimisant simplement d_1(U,V). Le désavantage est que la relation d'inégalité (3.16) ne tient plus et il n'est pas possible d'évaluer la borne supérieure sur l'erreur.

Résumons le travail accompli jusqu'à présent pour le cas unidimensionnel. Lorsque la fonction de coût est uniformément Lipschitz d'ordre 1, nous avons démontré que l'erreur de quantification était bornée supérieurement par une fonction composée de la constante de Lipschitz $\overline{L}_1$ et de la distance de Wasserstein d_1(U,V) associée à la métrique $\rho(u,v) = |u-v|$. La constante $\overline{L}_1$ ne dépend que du PS et peut parfois être difficile à calculer, mais elle ne nous empêche pas de procéder à la génération de scénarios. L'algorithme de Lloyd converge vers un minimum global de d_1(U,V) lorsque la distribution est continue et représente un bon choix de méthode pour trouver les scénarios. Finalement, si $z(x,\cdot)$ est uniformément Lipschitz d'ordre p, alors la quantification optimale revient à minimiser la distance de Wasserstein après la transformation de l'espace χ définie en (3.28) et (3.29).

3.3.2 Cas multidimensionnel

Nous devons maintenant généraliser les résultats de la section précédente au cas multidimensionnel. En une dimension, nous avons utilisé la distance de Wasserstein avec la métrique

$\rho(u,v) = |u-v|$. Pour les vecteurs aléatoires, nous utiliserons la métrique suivante :

$$\rho_2(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{D} a_i \ |u_i - v_i| \ , \tag{3.30}$$

où les $a_i,\ i = 1, ..., D$, sont des constantes positives. Notons que si $a_i = 1 \ \forall \ i$, alors ρ_2 coïncide avec la distance induite par la norme 1

$$\|\mathbf{u}-\mathbf{v}\|_1 = \sum_{i=1}^{D} |u_i - v_i| \ . \tag{3.31}$$

La métrique (3.31) est également appelée distance en valeur absolue ou distance de Manhattan. Dans le cas unidimensionnel ($D = 1$), $\|u-v\|_p = |u-v|$ peu importe le choix de p. L'avantage de la métrique (3.30) est qu'elle permet de pondérer la contribution de chaque dimension sur l'optimisation. En général, les variables aléatoires n'ont pas toutes la même influence sur le PS. Supposons par exemple que l'on veuille optimiser le rendement d'un portefeuille financier par l'optimisation stochastique. On considère que le portefeuille est composé d'actions dont la valeur fluctue de manière imprévisible et qui sont représentées mathématiquement par des variables aléatoires (continues). Or, si son capital financier est dominé par une certaine action, celle-ci aura une plus grande influence sur le PS. On accordera donc une plus grande importance à la distance entre cette variable et les scénarios qui la représentent, qui se traduira par un coefficient a_i plus élevé. On démontre maintenant qu'il existe une transformation de l'espace qui rend les distances de Wasserstein associées aux métriques (3.31) et (3.30) équivalentes. Cette transformation ne sera toutefois pas utilisée dans ce mémoire puisque les variables aléatoires ont environ toutes la même importance pour les deux PS considérés au dernier chapitre. Elle pourrait cependant être nécessaire pour des problèmes tels que celui du rendement de portefeuille énoncé ci-dessus, où les variables aléatoires n'ont pas toutes la même influence sur l'optimisation stochastique.

Proposition 3.4. *Soit d_1 et c_1 les distances de Wasserstein associées aux métriques $\rho_1(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{D} |u_i - v_i|$ et $\rho_2(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{D} a_i \ |u_i - v_i|$ respectivement, où les a_i sont des constantes positives. On définit l'application $\Gamma : \mathbb{R}^D \to \mathbb{R}^D$ par*

$$\Gamma(u_1, u_2, ..., u_D) = (a_1 u_1, a_2 u_2, ..., a_D u_D) \ . \tag{3.32}$$

Alors

$$d_1(\Gamma(\mathbf{U}), \Gamma(\mathbf{V})) = c_1(\mathbf{U}, \mathbf{V}) \ . \tag{3.33}$$

Preuve On a

$$\begin{aligned} d_1(\Gamma(\mathbf{U}),\Gamma(\mathbf{V})) &= \inf_{F_{\Gamma(\mathbf{U}),\Gamma(\mathbf{V})}} E\left[\rho_1(\mathbf{U},\mathbf{V})\right] \\ &= \inf_{F_{\Gamma(\mathbf{U}),\Gamma(\mathbf{V})}} E\left[\sum_{i=1}^{D} |a_iU_i - a_iV_i|\right] \\ &= \inf_{F_{\mathbf{U},\mathbf{V}}} E\left[\sum_{i=1}^{D} |a_iU_i - a_iV_i|\right] \qquad (3.34) \\ &= \inf_{F_{\mathbf{U},\mathbf{V}}} E\left[\sum_{i=1}^{D} a_i|U_i - V_i|\right] \\ &= \inf_{F_{\mathbf{U},\mathbf{V}}} E\left[\rho_2(\mathbf{U},\mathbf{V})\right] \\ &= c_1(\mathbf{U},\mathbf{V}) \ , \qquad (3.33) \end{aligned}$$

où l'égalité (3.34) provient du fait que Γ est une application bijective. □

Ainsi, on considérera à partir de maintenant que la transformation Γ a été appliquée si nécessaire *a priori* et que la distance de Wasserstein est associée à $\rho_1(\mathbf{u},\mathbf{v}) = \|\mathbf{u}-\mathbf{v}\|_1$. Le résultat suivant est présenté dans Pflug (2001) et généralise la proposition 3.1 au cas multidimensionnel.

Proposition 3.5. *Soit $\rho(\mathbf{u},\mathbf{v}) = \|\mathbf{u}-\mathbf{v}\|_1$ la métrique associée à la distance de Wasserstein et F_{U} la fonction de répartition de* U. *Si la fonction de coût satisfait la condition*

$$\inf\left\{L : |z(x,\mathbf{u}) - z(x,\mathbf{v})| \leq L \cdot \sum_{i=1}^{D} |u_i - v_i| \ \forall\ \mathbf{u} \in \Xi_{\boldsymbol{U}}, \mathbf{v} \in \Xi_{\boldsymbol{V}}\right\} \leq \tilde{L}_1 \quad \forall\ x \in X \ , \quad (3.35)$$

où $\Xi_{\boldsymbol{U}}$ et $\Xi_{\boldsymbol{V}}$ sont les supports des vecteurs $\mathbf{U}$ *et* $\mathbf{V}$ *respectivement et X est le domaine réalisable du PS, alors*

$$e(Z,\tilde{Z}) \leq 2\ \tilde{L}_1 \cdot d_1(\mathbf{U},\ \mathbf{V}) \qquad (3.36)$$

Preuve Une démonstration de cette proposition est présentée suite au Lemme 2 de Pflug (2001). La preuve peut également être faite de manière analogue à celle de la proposition 3.1 en posant $\psi(x,\mathbf{u}) = z(x,\mathbf{u})/\tilde{L}_1$.

Nous obtenons donc une borne supérieure similaire à celle obtenue en une dimension. Encore une fois, $\tilde{L}_1$ fait le lien entre le PS et $e(Z,\tilde{Z})$, tandis que $d_1(\mathbf{U},\ \mathbf{V})$ ne dépend que des distributions U et V. Si $z(x,u_i)$ est Lipshitz d'ordre $p_i \geq 1\ \forall\ i \in \{1,...,D\}$, alors la

transformation (3.28) se généralise au cas multidimensionnel par l'application $\chi_{p_i} : \mathbb{R}^D \to \mathbb{R}^D$ suivante (voir Pflug (2001)) :

$$\chi(u_1, u_2, ..., u_D) = (|u_1|^{p_1}\ sgn(u_1), |u_2|^{p_2}\ sgn(u_2), ..., |u_D|^{p_D}\ sgn(u_D))\ . \tag{3.37}$$

Nous allons donc considérer à partir de maintenant que la fonction de coût est uniformément Lipschitz d'ordre 1, ou que la transformation χ_p a été appliquée si nécessaire. Nous donnons maintenant la généralisation de la proposition 3.2 au cas multidimensionnel présentée dans Pflug (2001), qui n'a pas jugé nécessaire d'y introduire la preuve. Notre version de la démonstration est analogue à celle de la proposition 3.2 et est présentée par souci de clarté.

Proposition 3.6. *Soit* $\mathbf{U}$ *une distribution avec fonction de répartition* $F_{\mathbf{U}}$ *et* $\mathbf{V}$ *une distribution discrète de support fini qui prend les valeurs* $\mathbf{v}_j,\ j = 1, ..., K$*, avec probabilité* p_j*. Soit* $\rho(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{D} |u_i - v_i|$ *la métrique associée à la distance de Wasserstein. On définit les ensembles*

$$\Omega_j = \left\{ \mathbf{u} : \rho(\mathbf{u}, \mathbf{v}_j) \leq \min_k\ \rho(\mathbf{u}, \mathbf{v}_k) \right\}\ . \tag{3.38}$$

Les probabilités p_j *qui minimisent* $d_1(\mathbf{U},\mathbf{V})$ *sont*

$$p_j = \int_{\Omega_j} f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u}\ . \tag{3.39}$$

De plus, la distance de Wasserstein est donnée par

$$d_1(\mathbf{U},\mathbf{V}) = \sum_{j=1}^{K} \int_{\Omega_j} \sum_{i=1}^{D} |u_i - v_i|\ f_{\mathbf{U}}(\mathbf{u})\ d\mathbf{u}\ , \tag{3.40}$$

où $v_0 = -\infty$ *et* $v_{K+1} = \infty$*. Le supremum de* (3.13) *est atteint pour* $\psi(\mathbf{u}) = \min_{j\in}\ \rho(\mathbf{u}, \mathbf{v}_j)$*.*

Preuve À partir de la définition de la distance de Wasserstein, on a

$$\begin{aligned} d_1(\mathbf{U},\mathbf{V}) &= \min_{F_{\mathbf{U},\mathbf{V}}}\ E[\rho(\mathbf{U},\mathbf{V})] \\ &= \min_{F_{\mathbf{U},\mathbf{V}}}\ E\left[\sum_{i=1}^{D} |\mathbf{U}_i - \mathbf{V}_i|\right] \\ &= \min_{F_{\mathbf{U},\mathbf{V}}}\ \sum_{j=1}^{K} \int \sum_{i=1}^{D} |u_i - v_{ji}|\ f_{\mathbf{U}|\mathbf{V}}(\mathbf{u}|\mathbf{v}_j)\ d\mathbf{u}\ p_j \end{aligned} \tag{3.41}$$

Il est évident que $d_1(\mathbf{U}, \mathbf{V})$ sera minimal s'il est possible choisir une distribution conjointe qui associe chaque réalisation de $\mathbf{U}$ à la valeur $\mathbf{v}_j$ qui minimise $\rho(\mathbf{u},\mathbf{v}_j)$ avec probabilité 1,

i.e.

$$P\left[\mathbf{U} \in \Omega_j \mid \mathbf{V} = \mathbf{v}_j\right] = 1 \tag{3.42}$$

$$\text{et} \qquad P\left[\mathbf{V} = \mathbf{v}_j \mid \mathbf{U} \in \Omega_j\right] = 1 \ . \tag{3.43}$$

Par le théorème de Bayes, on obtient

$$\frac{P\left[\mathbf{V} = \mathbf{v}_j \mid \mathbf{U} \in \Omega_j\right] \ P\left[\mathbf{U} \in \Omega_j\right]}{p_j} = P\left[\mathbf{U} \in \Omega_j \mid \mathbf{V} = \mathbf{v}_j\right] = 1 \ ,$$

d'où

$$\begin{aligned} p_j &= P\left[\mathbf{U} \in \Omega_j\right] \\ &= \int_{\Omega_j} f_{\mathbf{U}}(\mathbf{u}) \ d\mathbf{u} \ . \end{aligned} \tag{3.19}$$

Notons que l'union des ensembles Ω_j couvre $\mathbb{R}^D$ et que pour une variable aléatoire continue $\mathbf{U}$, $P_{j\neq k}[\mathbf{U} \in \Omega_j \cap \mathbf{U} \in \Omega_k] = 0$. L'équation (3.41) devient alors

$$\begin{aligned} d_1(\mathbf{U}, \mathbf{V}) &= \sum_{j=1}^{K} \int_{\Omega_j} \sum_{i=1}^{D} |u_i - v_{ji}| \ f_{\mathbf{U}|\mathbf{V}}(\mathbf{u}|\mathbf{v}_j) \ d\mathbf{u} \ p_j \\ &= \sum_{j=1}^{K} \int_{\Omega_j} \sum_{i=1}^{D} |u_i - v_{ji}| \ f_{\mathbf{U},\mathbf{V}}(\mathbf{u}, \mathbf{v}_j) \ d\mathbf{u} \\ &= \sum_{j=1}^{K} \int_{\Omega_j} \sum_{i=1}^{D} |u_i - v_{ji}| \ f_{\mathbf{U}}(\mathbf{u}) \ d\mathbf{u} \ . \end{aligned} \tag{3.40}$$

Pour la dernière partie de la démonstration, on remplace $\psi(\mathbf{u}) = \min_j \rho(\mathbf{u}, \mathbf{v}_j)$ dans la formulation duale de la distance de Wasserstein donnée par (3.13).

$$\begin{aligned} d_1(\mathbf{U},\mathbf{V}) &= \left| \int \min_j \rho(\mathbf{u}, \mathbf{v}_j) \ dF_{\mathbf{U}}(\mathbf{u}) - \int \min_j \rho(\mathbf{v}, \mathbf{v}_j) \ dF_{\mathbf{V}}(\mathbf{v}) \right| \\ &= \left| \int \min_j \rho(\mathbf{u}, \mathbf{v}_j) \ dF_{\mathbf{U}}(\mathbf{u}) - \sum_{k=1}^{K} \min_j \rho(\mathbf{v}_k, \mathbf{v}_j) \ p_k \right| \\ &= \left| \int \min_j \rho(\mathbf{u}, \mathbf{v}_j) \ dF_{\mathbf{U}}(\mathbf{u}) \right| \\ &= \left| \int \min_j \ \sum_{i=1}^{D} |u_i - v_{ji}| \ dF_{\mathbf{U}}(\mathbf{u}) \right| \end{aligned} \tag{3.44}$$

$$= \sum_{j=1}^{K} \int_{\Omega_j} \sum_{i=1}^{D} |u_i - v_{ji}| \; f_{\mathbf{U}}(\mathbf{u}) \; d\mathbf{u} \; . \tag{3.40}$$

Par le théorème de Kantorovich-Rubinstein, $\psi(\mathbf{u})$ est donc la fonction qui maximise la forme duale de la distance de Wasserstein. □

On constate par la proposition ci-dessus que le calcul des scénarios v_j et de leur probabilité p_j est plus difficile pour le cas multidimensionnel. En effet, le calcul des intégrales de (3.39) et (3.40) exigent d'exprimer les frontières des ensembles Ω_j explicitement. Pour des valeurs v_j données, cette tâche est très ardue. Dans tous les cas, il est impossible de dériver (3.40) par rapport à v_j pour trouver le minimum de $d_1(\mathbf{U},\mathbf{V})$. Le résultat suivant est analogue à la proposition 3.3 et nous aidera à développer une méthode pour trouver les scénarios optimaux.

Proposition 3.7. *Soit* $\mathbf{U} = (\mathrm{U}_1, ..., \mathrm{U}_D)$ *une distribution avec fonction de répartition* $F_{\mathbf{U}}(\mathbf{u})$. *Supposons que les ensembles* $\Omega_j, \; j = 1, ..., K$, *forment une partition de* $\mathbb{R}^D$ *et que l'on associe une valeur* $\mathbf{v}_j \in \Omega_j$ *à chacun. Les valeurs* $\mathbf{v}_j = (v_{j1}, ..., v_{jD})$ *qui minimisent*

$$g(\mathbf{v}) = \sum_{j=1}^{K} \int_{\Omega_j} \sum_{i=1}^{n} |u_i - v_{ji}| \; f_{\boldsymbol{U}}(\boldsymbol{u}) \; d\boldsymbol{u} \tag{3.45}$$

correspondent aux médianes de chaque ensemble Ω_j, *i.e.*

$$P[U_i > v_{ji} | \, \boldsymbol{U} \in \Omega_j] = P[U_i < v_{ji} | \, \boldsymbol{U} \in \Omega_j] \quad i = 1, ..., D \; ; \; j = 1, ..., K \quad . \tag{3.46}$$

Preuve On dérive $g(\mathbf{v})$ par rapport à une variable quelconque $v_{j_0 i_0}$.

$$\begin{aligned}
0 &= \frac{\partial g(\mathbf{v}_1, ...\mathbf{v}_K)}{\partial v_{j_0 i_0}} \\
&= \frac{\partial}{\partial v_{j_0 i_0}} \int_{\Omega_{j_0}} \sum_{i=1}^{n} |u_i - v_{j_0 i}| \; f_{\mathbf{U}}(\mathbf{u}) \; d\mathbf{u} \\
&= \int_{\Omega_{j_0}} \frac{\partial}{\partial v_{j_0 i_0}} \sum_{i=1}^{n} |u_i - v_{j_0 i}| \; f_{\mathbf{U}}(\mathbf{u}) \; d\mathbf{u} \\
&= \int_{\Omega_{j_0}} \frac{\partial}{\partial v_{j_0 i_0}} |u_{i_0} - v_{j_0 i_0}| \; f_{\mathbf{U}}(\mathbf{u}) \; d\mathbf{u} \\
&= \int_{\Omega_{j_0}} sgn(u_{i_0} - v_{j_0 i_0}) \; f_{\mathbf{U}}(\mathbf{u}) \; d\mathbf{u} \\
&= \int_{\Omega_{j_0} : u_{i_0} > v_{j_0 i_0}} (1) \; f_{\mathbf{U}}(\mathbf{u}) \; d\mathbf{u} + \int_{\Omega_{j_0} : u_{i_0} < v_{j_0 i_0}} (-1) \; f_{\mathbf{U}}(\mathbf{u}) \; d\mathbf{u}
\end{aligned}$$

$$\begin{aligned}
\Longrightarrow \int_{\Omega_{j_0}:u_{i_0}>v_{j_0i_0}} f_{\mathbf{U}}(\mathbf{u})\ d\mathbf{u} &= \int_{\Omega_{j_0}:u_{i_0}<v_{j_0i_0}} f_{\mathbf{U}}(\mathbf{u})\ d\mathbf{u} \\
\int_{\Omega_{j_0}:u_{i_0}>v_{j_0i_0}} \frac{f_{\mathbf{U}}(\mathbf{u})}{P[\mathbf{U}\in\Omega_{j_0}]}\ d\mathbf{u} &= \int_{\Omega_{j_0}:u_{i_0}<v_{j_0i_0}} \frac{f_{\mathbf{U}}(\mathbf{u})}{P[\mathbf{U}\in\Omega_{j_0}]} f_{\mathbf{U}}(\mathbf{u})\ d\mathbf{u} \\
P[U_{i_0} > v_{j_0i_0}|\mathbf{U}\in\Omega_{j_0}] &= P[U_{i_0} < v_{j_0i_0}|\mathbf{U}\in\Omega_{j_0}] \qquad (3.47)
\end{aligned}$$

On pourrait faire de même pour tous les indices j et i ; les médianes correspondent donc au minimum de $g(\mathbf{v})$. □

En une dimension, on entend par médiane d'une variable aléatoire U la valeur ν pour laquelle $P[U > \nu] = P[U < \nu]$. La définition de médiane peut toutefois varier selon les auteurs dans le cas multidimensionnel. Nous supposerons dans ce mémoire que la médiane $\boldsymbol{\nu}$ d'un vecteur aléatoire $\mathbf{U}$ est définie par (3.46). C'est-à-dire que chaque composante de $\boldsymbol{\nu}$ correspond à la médiane unidimensionnelle des distributions marginales de $\mathbf{U}$.

Une proposition analogue à celle ci-dessus avait inspiré l'algorithme de Lloyd en une dimension. Cet algorithme se généralise facilement au cas multidimensionnel, mais ne garantit alors que la convergence vers un minimum local de $d_1(\mathbf{U}, \mathbf{V})$. De plus, l'intégrale multidimensionnelle sur les sous-espaces Ω_j est loin d'être évidente. Nous verrons cependant que la minimisation de la distance de Wasserstein est intimement liée au problème des k-médianes du domaine de la partition de données (*clustering analysis*), pour lequel l'algorithme de Lloyd n'est qu'une méthode de résolution parmi tant d'autres. Nous chercherons donc dans le prochain chapitre des méthodes pour minimiser $d_1(\mathbf{U,V})$ inspirées d'algorithmes du problème des k-médianes en espérant qu'ils nous permettront d'atteindre des solutions plus près du minimum global.

CHAPITRE 4

PARTITIONNEMENT DE DONNÉES

4.1 Introduction

Le partitionnement de données est une branche des mathématiques dont l'objectif est de classer les données selon des mesures de similarité (resp. dissimilarité) entre les éléments d'un même groupe (resp. de groupes différents). Il fait partie du domaine de la classification non supervisée puisque l'on ne connaît pas les groupes auxquels appartiennent les données *a priori*. Il existe plusieurs types de problèmes de partitionnement de données définis selon différents critères de similarité tels que la distance par rapport à un élément central, le maximum de vraisemblance d'une distribution ou la densité des éléments du groupe, et il existe une grande variété d'algorithmes de résolution dans la littérature pour chacun d'entre eux (Everitt *et al.* (2011)). Dans notre cas, on s'intéressera au problème des k-médianes défini ci-dessous, puisqu'on verra sous peu qu'il est équivalent au problème de minimisation de la distance de Wasserstein lorsque le nombre de données tend vers l'infini.

Définition 4.1. *Soit* $\mathbf{U}$ *un ensemble de* N *données,* $\mathbf{u}_n \in \mathbb{R}^D,\ n = 1, ..., N$*, et une métrique* $\rho(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{D} |u_i - v_i|$ *:* $\mathbb{R}^D \times \mathbb{R}^D \to [0, \infty)$*. Le* **problème des k-médianes** *consiste à chercher des valeurs* $\mathbf{v}_j \in \mathbb{R}^D,\ j = 1, ..., K$*, appelées* **centres** *qui minimisent*

$$z(\mathbf{v}) = \sum_{n=1}^{N} \min_j \rho(\mathbf{u}_n, \mathbf{v}_j) \ . \tag{4.1}$$

Si l'on choisit plutôt le carré de la distance euclidienne $\rho(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{D} |u_i - v_i|^2$, alors on obtient le problème des *k-moyennes*. On dit que les centres v_j induisent une partition de l'espace, appelée partition de Voronoï, en sous-ensembles

$$\Omega_j = \left\{ \mathbf{u} : \rho(\mathbf{u}, \mathbf{v}_j) \leq \min_k \rho(\mathbf{u}, \mathbf{v}_k) \right\} \ , \tag{4.2}$$

qui correspondent aux sous-espaces que l'on avait définis au chapitre précédent (voir définition (3.38)). Les partitions de Voronoï se visualisent particulièrement bien en 2 dimensions et permettent de vérifier qualitativement la qualité des centres obtenus (voir figure 4.1). L'union des Ω_j couvre $\mathbb{R}^D$ en entier et seuls les points sur les frontières (représentés par les segments de droite sur la figure (4.1)) appartiennent à deux sous-ensembles à la fois.

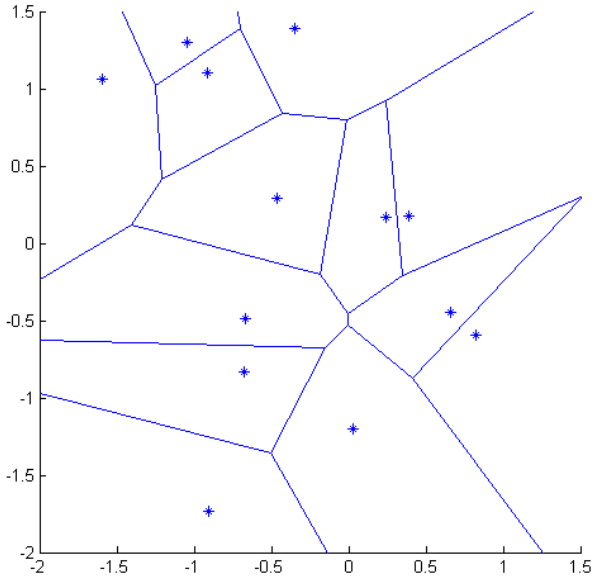

Figure 4.1 Diagramme de Voronoï pour 13 centres avec la distance euclidienne.

La différence entre la minimisation de la distance de Wasserstein par un ensemble de scénarios $\mathbf{v}_j$ et le problème des k-médianes provient du fait que $\mathbf{U}$ est une distribution aléatoire dans le premier cas et un ensemble de données dans le deuxième. Or, on constate à partir de (3.44) que la distance de Wasserstein peut également être réécrite comme

$$d_1(\mathbf{U},\mathbf{V}) = E\left[\min_j \rho(\mathbf{U}, \mathbf{v}_j)\right] \ . \tag{4.3}$$

De plus, si l'on suppose que les données $\mathbf{u}_n$ correspondent à des réalisations d'un vecteur aléatoire $\mathbf{U}$, alors par la loi forte des grands nombres,

$$\lim_{N\to\infty} \sum_{n=1}^{N} \min_j \rho(\mathbf{u}_n, \mathbf{v}_j) = E\left[\min_j \rho(\mathbf{U}, \mathbf{v}_j)\right] \ . \tag{4.4}$$

La minimisation de $d_1(\mathbf{U},\mathbf{V})$ est donc équivalente au problème des k-médianes lorsque les données correspondent à des échantillons de $\mathbf{U}$ et que leur quantité tend vers l'infini.

La quantité de solutions possibles pour le problème des k-médianes est de l'ordre du nombre

de Stirling de seconde espèce

$$S(N,K) = \frac{1}{K!}\sum_{j=1}^{K}(-1)^{K-j}\binom{K}{j}j^N \ , \tag{4.5}$$

qui représente le nombre de manières de partitionner N éléments en K sous-ensembles non vides (Graham *et al.* (1994)). En réalité, la quantité de solutions possibles est inférieure à $S(N,K)$, puisque certaines partitions sont impossibles pour une mesure de similarité donnée. Par exemple, si l'on souhaite représenter les 4 données $\mathbf{u}_1 = (1,1),\ \mathbf{u}_2 = (2,\ \frac{3}{2}),\ \mathbf{u}_3 = (3,1)$ et $\mathbf{u}_4 = (2,8)$ par 2 scénarios, alors la seule partition possible est représentée sur la figure 4.2, où les astérisques représentent les médianes de chaque groupe. Notons qu'une condition nécessaire pour former des groupes de données est que ceux-ci soient linéairement séparables deux à deux. Cette condition n'est toutefois pas suffisante. Par exemple, les sous-ensembles $\{\mathbf{u}_1, \mathbf{u}_4\}$ et $\{\mathbf{u}_2, \mathbf{u}_3\}$ de la figure 4.2 sont linéairement séparables, mais ne constituent pas une solution réalisable puisque $\mathbf{u}_1$ est plus près de la médiane de $\{\mathbf{u}_2,\ \mathbf{u}_3\}$ que de sa médiane avec $\mathbf{u}_4$. La partition de Voronoï des centres avec la distance L_1 est également représentée, où la distance en valeur absolue cause la division particulière de l'espace sur la figure 4.2. Malgré tout, le nombre de solutions possibles en dimension supérieure reste énorme et il n'existe pas de méthode simple pour les identifier. C'est pourquoi il est si difficile de trouver le minimum global au problème des k-médianes. De plus, même s'il est atteint, il n'est généralement pas possible de vérifier qu'il s'agit effectivement de la solution optimale.

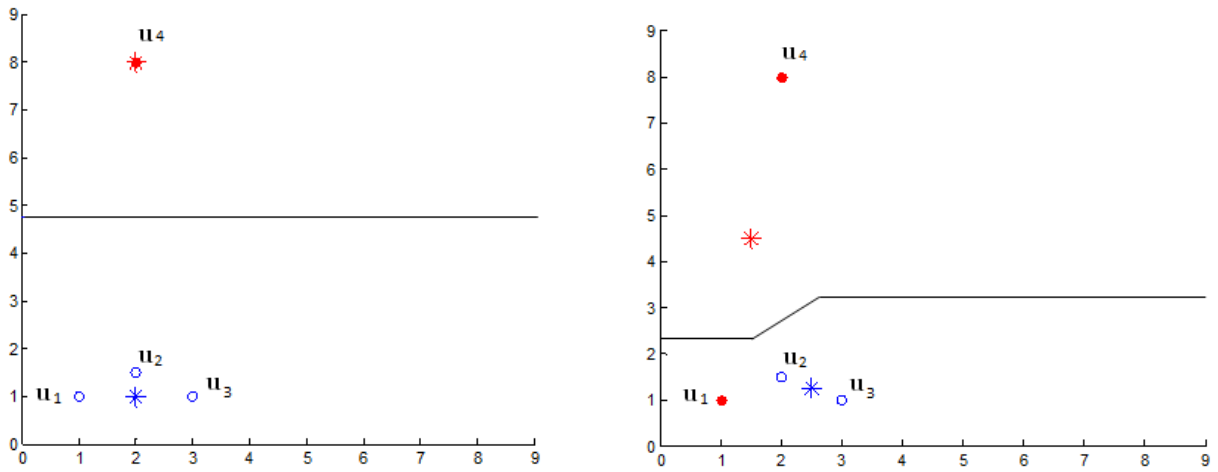

Figure 4.2 Partitions possible (gauche) et impossible (droite)

En une dimension, si $u_1 < u_2 < u_3$, alors il est impossible de former deux sous-ensembles Ω_1 et Ω_2 tels que $\{u_1, u_3\} \subset \Omega_1$ et $u_2 \in \Omega_2$. La quantité de K ensembles non vides que l'on peut former à partir de N données est donc bornée par le nombre de solutions entières positives de l'équation

$$x_1 + x_2 + ... + x_K = N \qquad x_i \in \mathbb{N}^*,\ i = 1, ..., K\ , \tag{4.6}$$

qui est égal à $\binom{N-1}{K-1}$. Cette quantité est bien inférieure au nombre de Stirling de seconde espèce. De plus, comme pour le cas multidimensionnel, ces partitions ne sont pas nécessairement toutes réalisables. Cependant, le nombre de sous-ensembles possibles reste énorme et il n'est généralement pas possible de tous les considérer. Ainsi, même en une dimension, il est difficile de trouver la solution optimale au problème des k-médianes sur un ensemble de N données.

Il importe de noter qu'en réalité, on ne connaît jamais exactement la distribution du vecteur aléatoire présent dans le problème. Celle-ci est normalement estimée à partir d'un ensemble de données statistiques. Or, il peut parfois être difficile d'estimer la distribution d'un vecteur aléatoire avec confiance à partir des données, particulièrement lorsque celles-ci sont en quantité insuffisante ou que la loi de probabilité comporte plusieurs paramètres. Afin d'estimer une loi de probabilité, on commence généralement par émettre une hypothèse quant à la nature de la distribution aléatoire ayant généré les données. Les paramètres de celles-ci sont ensuite estimés et un test statistique est utilisé pour vérifier la validité de la distribution obtenue. Il existe une très grande variété de tests statistiques. Nous suggérons l'usage du test de Pearson (1900), qui évalue la probabilité des écarts observés entre les données et la distribution estimée et qui est couramment utilisé dans la littérature. Si l'écart observé est peu probable, alors l'hypothèse selon laquelle les données proviennent de la loi aléatoire estimée est rejetée. Il nous semble raisonnable d'utiliser le test de Pearson avec un seuil de signification $\alpha = 0,05$, c'est-à-dire que si la distribution estimée est exacte, alors elle sera rejetée dans 5 % des cas. Une plus petite valeur de α diminuerait les chances de faire une telle erreur (de type I), mais augmente celles d'accepter la distribution alors qu'elle est fausse (erreur de type II).

Nous proposons deux façons distinctes de générer les scénarios. Si le test de Pearson ne permet pas de rejeter la distribution estimée, on supposera que les données proviennent effectivement de celle-ci. On peut alors échantillonner un très grand nombre de valeurs à partir de la distribution qui serviront ensuite d'entrées pour l'algorithme des k-médianes. L'avantage

de cette procédure est que l'on peut tirer autant d'échantillons que l'on désire pour chercher les centres $\mathbf{v}_j$. Au contraire, si le test de Pearson échoue, alors nous suggérons de chercher les valeurs optimales $\mathbf{v}_j^*$ en utilisant directement un algorithme du problème des k-médianes à partir de l'ensemble de données à notre disposition. L'avantage dans ce cas est que l'on peut procéder à la génération de scénarios sans avoir à estimer la distribution du vecteur aléatoire. En effet, la distribution estimée est une approximation du phénomène probabiliste réel tandis que les données proviennent d'un historique et collent forcément à la réalité. Nous allons donc présenter dans les prochaines sections des algorithmes du problème des k-médianes pour chacune de ces procédures.

Notons qu'il est toujours possible de trouver une distribution par régression qui correspond parfaitement à un ensemble de données quelconque et qui ne sera rejetée par aucun test statistique. Il faut cependant rester prudent afin d'éviter le «surajustement» (*overfitting*) de la distribution aux données et ainsi obtenir un modèle manquant de généralité. Tout comme une quantité de données insuffisante ou une valeur trop élevée du seuil de signification α, le surajustement augmente la probabilité β de commettre une erreur de type II. Idéalement, on voudrait procéder à l'estimation de la distribution lorsque les valeurs de α et β sont assez faibles, mais ceci peut être difficile à accomplir en réalité. L'élaboration d'un critère ou d'un test statistique permettant de déterminer si l'on devrait procéder à l'estimation de la distribution dépasse le cadre de ce mémoire et est laissée ouverte à de futures recherches.

4.2 Distribution connue

Si l'on choisit d'estimer la distribution du vecteur aléatoire, on peut alors tirer autant d'échantillons qu'on le désire à partir de la loi estimée. Les échantillons de la distribution sont alors équivalents à des données, qui servent d'entrées pour les algorithmes de partitionnement. En une dimension, on utilisera l'algorithme de Lloyd continu puisque nous avons vu au chapitre précédent qu'il converge vers le minimum global de la distance de Wasserstein. Cependant, il est beaucoup plus difficile à implémenter dans le cas multidimensionnel. Nous présentons dans cette section un algorithme basé sur la méthode de descente du gradient stochastique et conçu pour discrétiser des lois aléatoires multinomiales continues. Nous supposons dans ce qui suit que la distribution du vecteur aléatoire a été estimée *a priori*.

Définition 4.2. *Soit* $\mathbf{U}$ *un vecteur aléatoire avec fonction de répartition* $F_{\mathbf{U}}(\mathbf{u})$ *et* $H : \mathbb{R}^D \to \mathbb{R}$. *Supposons que l'on cherche à minimiser*

$$H(\mathbf{v}) = \int h(\mathbf{v}, \mathbf{u}) \, dF_{\mathbf{U}}(\mathbf{u}) \tag{4.7}$$

et que $\nabla h(\mathbf{v}, \mathbf{u})$ *existe* $\forall\ \mathbf{v}, \mathbf{u}$. *La méthode de* **descente du gradient stochastique** *consiste à ajuster les valeurs de* $\mathbf{v}$ *séquentiellement à partir d'une valeur initiale* $\mathbf{v}^0$ *quelconque par la formule suivante :*

$$\mathbf{v}^{n+1} = \mathbf{v}^n - \gamma_{n+1}\ \nabla h(\mathbf{v}^n, \mathbf{u}_n)\ , \tag{4.8}$$

où $\gamma > 0$ *est appelé paramètre de saut,* $\mathbf{v}^n$ *est la valeur de* $\mathbf{v}$ *à l'itération* n *et les valeurs de* $\mathbf{u}_n$ *correspondent à des réalisations de* $\mathbf{U}$.

Il est possible de démontrer sous diverses conditions que

$$P\left[\lim_{n\to\infty} \|\mathbf{v}^n - \mathbf{v}^*\| > \epsilon\right] = 0\ , \tag{4.9}$$

pour toute constante $\epsilon > 0$ arbitrairement petite, où $\mathbf{v}^n$ est la suite de valeurs générées par la descente du gradient stochastique avec la norme $\|\cdot\|$ et $\mathbf{v}^*$ est un point stationnaire de H (voir par exemple Pages et Printems (2003) ou Cardot *et al.* (2013)). Malheureusement, les preuves de convergence rencontrées impliquent certaines conditions qui ne sont pas satisfaites pour la norme L_1. On se basera néanmoins sur la descente du gradient stochastique pour développer un algorithme de génération de scénarios adapté à notre problème.

Dans notre cas, on cherche les valeurs $\mathbf{v}_j$ qui minimisent la distance de Wasserstein. À partir de (3.40), on constate que ceci revient à minimiser

$$H(\mathbf{v}_j) = \int_{\Omega_j} \sum_{i=1}^{D} |u_i - v_{ji}|\ f_{\mathbf{U}}(\mathbf{u})\ d\mathbf{u}\ . \tag{4.10}$$

On a donc

$$h(\mathbf{v}_j, \mathbf{u}) = \sum_{i=1}^{n} |u_i - v_{ji}| = \rho(\mathbf{u}, \mathbf{v}_j) \tag{4.11}$$

$$\text{et} \qquad \frac{\partial h(\mathbf{v}_j, \mathbf{u})}{\partial v_{ji}} = -sgn\ (u_i - v_{ji})\ . \tag{4.12}$$

Il ne nous reste plus qu'à déterminer le paramètre de saut. Cette partie est la plus difficile, car la descente du gradient stochastique est très sensible au choix de γ_n. Pour des raisons de convergence, le paramètre de saut est normalement choisi de manière à satisfaire les conditions suivantes :

$$\sum_{n=1}^{\infty} \gamma_n = \infty \quad et \quad \sum_{n=1}^{\infty} \gamma_n^2 < \infty\ . \tag{4.13}$$

Une forme standard pour ce paramètre est

$$\gamma_n = \frac{b}{m + n^\alpha} \ , \tag{4.14}$$

où $b > 0$ et $m \geq 0$ sont des constantes et $\alpha \in (1/2, 1)$ (voir par exemple Pages et Printems (2003), Cardot *et al.* (2013) ou Moulines et Bach (2011)). Notons que ce choix de paramètre satisfait les conditions (4.13). Nous reviendrons sur l'ajustement des valeurs b, m et α, mais présentons d'abord un algorithme basé sur la méthode de descente du gradient stochastique connu sous le nom de quantification vectorielle par apprentissage compétitif, qui nous permettra de générer les scénarios pour l'optimisation stochastique.

On introduit la fonction suivante qui sera utilisée dans l'algorithme de quantification vectorielle :

$$\mathbb{I}_j(\mathbf{u}) = \begin{cases} 1 & si\ h(\mathbf{v}_j^n, \mathbf{u}) \leq \min\limits_l h(\mathbf{v}_l^n, \mathbf{u}) \\ 0 & sinon\ . \end{cases} \tag{4.15}$$

Quantification vectorielle par apprentissage compétitif (QVAC)

1. (*Initialisation*) On tire des échantillons $\mathbf{u}_j,\ j = 1, ..., K$, à partir de la distribution $\mathbf{U}$. On pose $\mathbf{v}_j^0 = \mathbf{u}_j$ et $p_j^0 = 0\ \forall\ j$. On pose également $n = 0$ et on choisit un nombre d'itérations $N \in \mathbb{N}$.
2. (*Phase compétitive*) On tire un nouvel échantillon $\mathbf{u}$ de $\mathbf{U}$ et on détermine un indice j tel que $h(\mathbf{v}_j^n, \mathbf{u}) \leq \min\limits_l h(\mathbf{v}_l^n, \mathbf{u})$.
3. (*Phase d'apprentissage*) On pose

$$v_{ij}^{n+1} = v_{ij}^n + \mathbb{I}_j(\mathbf{u})\ (\gamma_{n+1}\ sgn(u_i - v_{ij})) \tag{4.16}$$

$$\text{et} \qquad p_j^{s+1} = p_j^s + \mathbb{I}_j(\mathbf{u}) \tag{4.17}$$

   où γ_{n+1} est le paramètre de saut de la méthode de descente du gradient stochastique.
4. (*Terminaison*) Si $n < N$, alors $n \leftarrow n + 1$ et on retourne à la deuxième étape. Sinon, les scénarios correspondent aux valeurs $\mathbf{v}_j^N$ avec probabilités respectives p_j^N/N.

L'algorithme ci-dessus nous permet non seulement de trouver les médianes des sous-ensembles Ω_j, mais aussi les probabilités p_j associées à chaque ensemble Ω_j. Il serait possible de démontrer que les valeurs de p_j convergent effectivement vers $P[\mathbf{U} \in \Omega_j]$. Par les propositions 3.6 et 3.7, les valeurs obtenues par la quantification vectorielle par apprentissage compétitif minimisent la distance de Wasserstein.

Une autre façon de calculer les probabilités consiste à échantillonner de nouveau la distribution après avoir trouvé les scénarios. Ce procédé est toutefois un peu plus laborieux. Nous utiliserons donc la méthode décrite ci-haut pour assigner les probabilités aux scénarios. Il existe également une alternative à l'initialisation ci-dessus dont nous nous servirons à quelques reprises au chapitre suivant. Plutôt que de prendre des échantillons aléatoires comme scénarios initiaux, ces derniers peuvent aussi correspondre aux scénarios obtenus par une implémentation antérieure de la QVAC. Cette procédure est équivalente à exécuter la QVAC deux fois de suite, où les scénarios initiaux de la 2^e implémentation correspondent à ceux obtenus par la 1^{re}.

On se concentre maintenant sur l'ajustement du paramètre de saut. Comme nous l'avons déjà mentionné, la convergence de la méthode du gradient stochastique (et celle de la QVAC par le fait même) est très sensible aux valeurs de b, c et α. En pratique, il est donc préférable de déterminer ces valeurs empiriquement selon la dimension, le nombre de médianes et la distribution de $\mathbf{U}$ pour le problème considéré. Cependant, étant donné que nous effectuerons plusieurs simulations au prochain chapitre à partir d'instances variées, il sera plus commode de trouver une forme générale du paramètre de saut qui donne des résultats acceptables pour toutes les instances. On supposera pour commencer que $m > 0$, ce qui nous permet d'écrire

$$\gamma_n = \beta \, \frac{m}{m + n^\alpha} \, , \tag{4.18}$$

où $\beta > 0$ est une nouvelle constante. Puisque l'on veut effectuer une grande quantité de tests numériques, on accélérera la convergence de la QVAC en prenant $\alpha = 1$. L'ajustement du paramètre de saut repose donc sur la recherche des valeurs β et m qui permettent d'obtenir les meilleures discrétisations. La valeur de β correspond à la grandeur du pas initial (lorsque $n = 0$) tandis que m contrôle le taux de décroissance de γ_s. Plus m est petit, plus le paramètre de saut décroîtra rapidement.

En général, l'ajustement du paramètre de saut constitue un problème difficile et dépend du vecteur aléatoire que l'on souhaite quantifier. Cependant, puisque diverses distributions seront testées au prochain chapitre, il serait assez laborieux d'ajuster le paramètre de saut pour chacune d'entre elles. Nous présenterons donc une heuristique qui permet d'obtenir des valeurs satisfaisantes de β et m en supposant que les distributions marginales sont non corrélées et suivent une loi normale standard $\mathcal{N}(0, 1)$. L'ajustement de γ_n ne sera donc pas optimal pour les autres distributions, mais nous verrons au prochain chapitre que le paramètre

de saut obtenu avec la loi normale standard entraîne également des gains importants (par rapport à l'échantillonnage pur) pour plusieurs autres distributions. La distribution étant fixée, les valeurs optimales de β et m ne dépendent plus que du nombre de médianes K et de la dimension que l'on notera D. On cherche donc une forme de γ_n qui s'ajuste quelque peu au nombre de scénarios et à la dimension du problème :

$$\gamma_n(K, D) = \beta(K, D)\ \frac{m(K, D)}{m(K, D) + n}\ . \tag{4.19}$$

Nous présentons maintenant notre heuristique qui permet d'ajuster les paramètres itérativement en fonction de trois caractéristiques du problème : la dimension, le nombre de centres et les variances marginales. L'influence de chacune de ces caractéristiques est analysée tour à tour dans les sous-sections qui suivent. Nous avons utilisé des distributions marginales $\mathcal{N}(0, 1)$, mais la même procédure peut également s'appliquer aux autres lois aléatoires. Les tests des prochaines sous-sections sont tous effectués à l'aide de la QVAC avec $N = 500\ 000$ itérations.

Dimension

On commence par poser $\beta(K, D) = 1$ et évaluer l'influence de la dimension sur l'ajustement de $m(K, D)$. Nous avons supposé que $m(K, D) = K \cdot m_D$, où $m_D \in \mathbb{R}$ est un paramètre qui dépend de la dimension. On obtient donc

$$\gamma_n(K, D) = 1 \times \frac{K \cdot m_D}{K \cdot m_D + n}\ . \tag{4.20}$$

Nous avons ensuite implémenté l'algorithme de QVAC avec différentes valeurs de m_D et calculé la distance de Wasserstein des solutions obtenues. Les courbes de $d_1(\mathbf{U}, \mathbf{V})$ en fonction de m_D sont présentées à la figure 4.3 pour $D = 2,\ 5,\ 10$ et 20. On ne considère pas le cas unidimensionnel ici puisque l'on a établi que la méthode de Lloyd serait alors utilisée. On constate que l'on obtient de bons résultats avec $m_D \approx 7$ peu importe la dimension et que les solutions obtenues sont très stables autour de cette valeur. On se satisfera donc de cette valeur de m_D pour l'instant.

On évalue maintenant le comportement de $\beta(K, D)$ en fonction de la dimension en posant $\beta(K, D) = \beta_D$. Suite aux résultats ci-dessus, on a

$$\gamma_n(K, D) = \beta_D \times \frac{7K}{7K + n} \qquad \textit{pour } D = 2,\ 3, ... \tag{4.21}$$

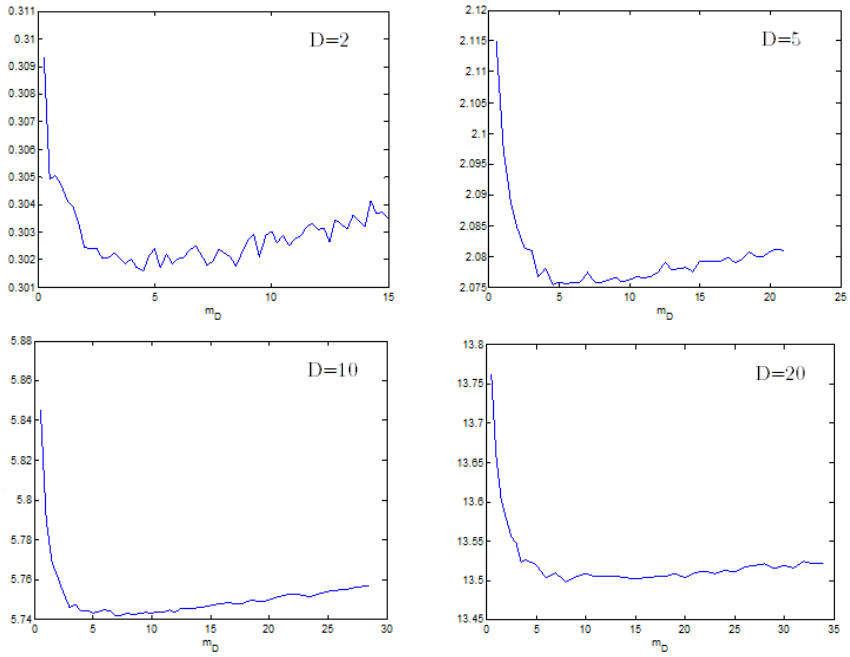

Figure 4.3 Distance de Wasserstein en fonction du paramètre m_D pour les dimensions 2, 5, 10 et 20.

De manière analogue à ce qui a été fait ci-haut, les courbes de la distance de Wasserstein en fonction du paramètre β_D pour $D = 2,\ 5,\ 10$ et 20 se retrouvent sur la figure 4.4. On constate que l'on devrait probablement choisir une valeur de β_D légèrement inférieure à 1. Sauf en 2 dimensions, les distances de Wasserstein les plus basses sont obtenues lorsque $\beta_D \approx 0,8$. Même lorsque $D = 2$, la différence entre les solutions minimales et celle avec $\beta_D = 0,8$ est négligeable. On prendra donc cette valeur de β pour toutes les dimensions.

Nombre de centres

On considère maintenant l'influence du nombre de centres sur l'ajustement du paramètre de saut. On a choisi d'effectuer les tests avec $D = 2$ pour accélérer la convergence. On cherche les valeurs m_K du paramètre de saut de la forme

$$\gamma_n = 0.8 \times \frac{7\ m_K}{7\ m_K + n} \qquad pour\ D = 2,\ 3, ... \tag{4.22}$$

qui minimise la distance de Wasserstein. Les résultats pour $K = 5,\ 50,\ 100$ et 200 en 2 dimensions sont présentés à la figure 4.5. Puisqu'on a utilisé $K = 50$ pour l'ajustement du facteur de saut en fonction de la dimension et que l'on avait supposé $m_K = K$, on s'attend à obtenir les meilleurs résultats pour 50 scénarios lorsque $m_K = 50$. On constate que n'importe

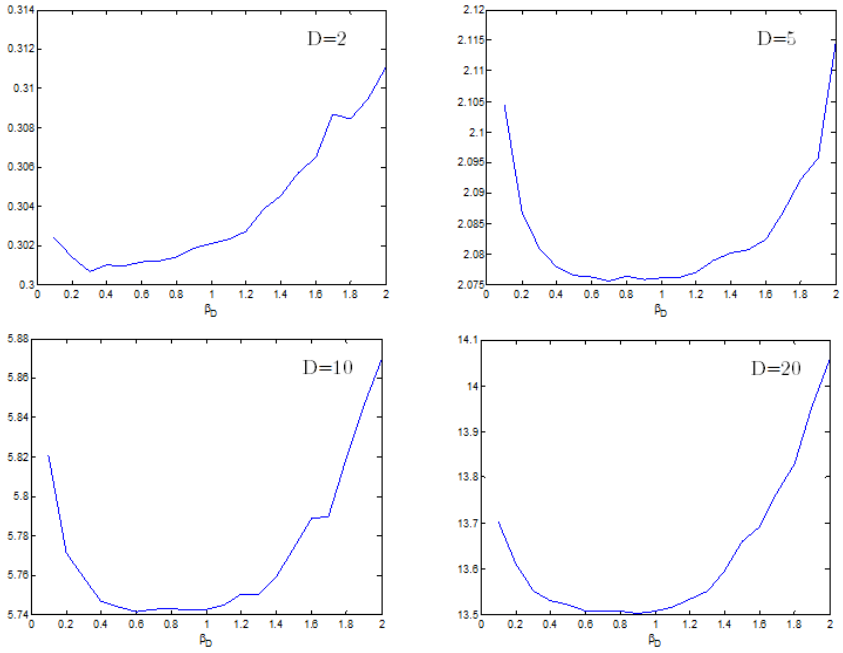

Figure 4.4 Distance de Wasserstein en fonction du paramètre β_D pour les dimensions 2, 5, 10 et 20.

quelle valeur de m_K entre 20 et 60 donne de bons résultats lorsque $D = 2$ et $K = 50$. Pour le reste des instances avec $D = 2$, on trouve cependant que m_K devrait se situer autour de 25 indépendamment de K.

On pourrait démontrer de manière analogue aux tests précédents que $\beta = 0,8$ est un choix approprié, peu importe le nombre de centres. Le paramètre de saut pour la QVAC aura donc la forme suivante :

$$\gamma_n = 0,8 \times \frac{7 \cdot 25}{7 \cdot 25 + n} = 0,8 \times \frac{175}{175 + n} \qquad pour\ D = 2,\ 3, ... \tag{4.23}$$

Il est assez surprenant de constater que pour $D \geq 2$, l'ajustement de γ_n dépend peu de la dimension et du nombre de centres. Nous avons donc choisi le paramètre de saut indépendant de ces facteurs. Nous ne tenterons pas une explication simpliste de ce phénomène assez complexe. Notons cependant qu'un nombre élevé de médianes diminuent les chances de chacune d'entre elles d'être ajustée à chaque itération, mais augmente leur chance de se trouver près de leur valeur optimale $\mathbf{v}_j^*$.

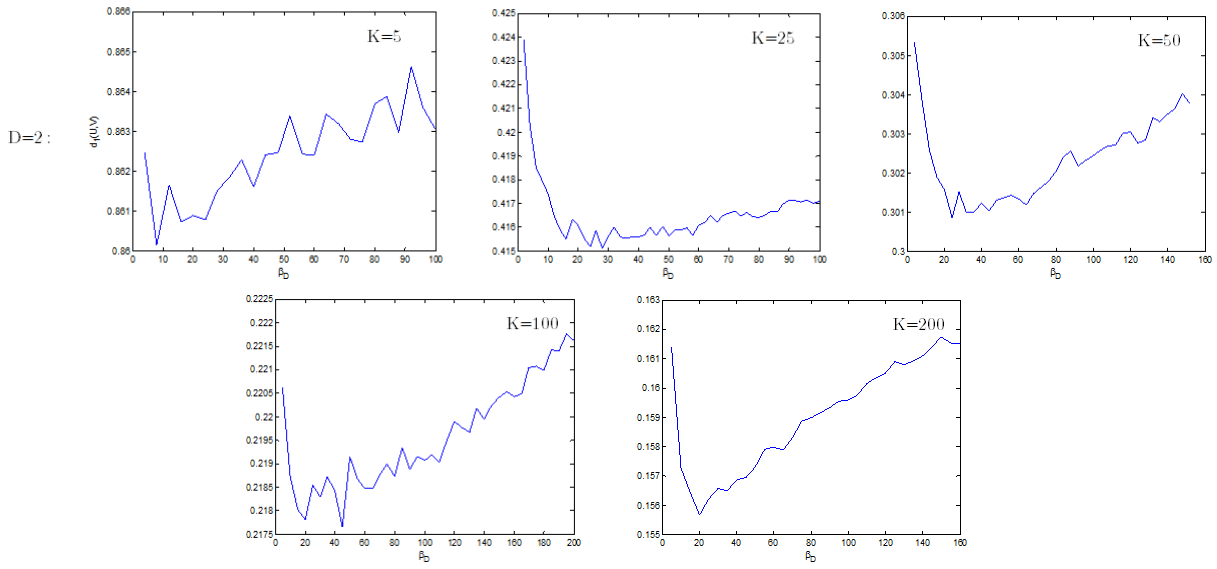

Figure 4.5 Distance de Wasserstein en fonction du paramètre m_K pour $K = 5$, 25, 50, 100 et 200.

Variances marginales

Le dernier aspect que nous analyserons est l'influence des variances marginales sur le paramètre de saut. Prenons par exemple une variable aléatoire U suivant une distribution $\mathcal{N}(\mu, \sigma^2)$. On a alors

$$\begin{aligned}
E\left[|U-\mu|\right] &= \int_{-\infty}^{\infty} |u-\mu| \ \frac{1}{\sqrt{2\pi}\sigma} \ e^{\frac{-(u-\mu)^2}{2\sigma^2}} \ du \\
&= \int_{-\infty}^{\mu} (\mu-u) \ \frac{1}{\sqrt{2\pi}\sigma} \ e^{\frac{-(u-\mu)^2}{2\sigma^2}} \ du + \int_{\mu}^{\infty} (u-\mu) \frac{1}{\sqrt{2\pi}\sigma} \ e^{\frac{-(u-\mu)^2}{2\sigma^2}} \ du \\
&= -\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\sigma} \ e^{\frac{-v}{2\sigma^2}} \ \frac{dv}{2} + \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \ e^{\frac{-v}{2\sigma^2}} \ \frac{dv}{2} \\
&= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \ e^{\frac{-v}{2\sigma^2}} \ dv \\
&= \left[\frac{1}{\sqrt{2\pi}\sigma} \ e^{\frac{-v}{2\sigma^2}} \ (-2\sigma^2) \right]_0^{\infty} \\
&= \sqrt{\frac{2}{\pi}} \ \sigma \ . \qquad (4.24)
\end{aligned}$$

En comparaison avec une $\mathcal{N}(\mu, 1)$, l'espace de probabilité d'une variable aléatoire $\mathcal{N}(\mu, \sigma^2)$ est donc dilaté d'un facteur σ par rapport à sa moyenne. Par conséquent, le paramètre de

saut doit être ajusté d'un facteur σ_i :

$$\gamma_n = 0,8\ \sigma_i \times \frac{175}{175+n} \qquad pour\ D = 2,\ 3, ... \tag{4.25}$$

où σ_i est l'écart-type de la distribution marginale $\mathrm{U}_i,\ i = 1, ..., D$. On constate également que le paramètre de saut ne dépend pas de la moyenne μ, qui n'influence que le centrage de la fonction de densité.

L'inégalité (3.16) nous a conduits à considérer la distance induite par la norme L_1 et la métrique $\rho(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{D} |u_i - v_i|$. Or, rien ne nous empêche de considérer une autre mesure de distance. Par exemple, on peut choisir $\rho(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{D} |u_d - v_d|^2$ si l'on souhaite pénaliser plus fortement les grands écarts entre U et les scénarios. Le problème des k-médianes devient alors celui des k-moyennes. Le seul désavantage est que l'on ne pourra pas évaluer la borne supérieure sur l'erreur d'approximation donnée par (3.16), mais on verra que celle-ci est souvent beaucoup plus élevée que l'erreur réellement commise.

Les algorithmes vus jusqu'à présent se généralisent tous facilement à cette nouvelle distance. Pour l'algorithme de Lloyd continu, il suffit de calculer la moyenne plutôt que la médiane en 3e étape. Pour la QVAC, on a $h(\mathbf{v}_j, \mathbf{u}) = \rho(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{D} |u_d - v_d|^2$ et on trouve

$$\frac{\partial h(\mathbf{v}_j, \mathbf{u})}{\partial v_{ji}} = -2\ (u_i - v_{ji})\ .$$

L'avantage de considérer la norme L_2 est que le saut de chaque itération de la QVAC est proportionnel à la différence entre u_i et v_{ji}. Pour la norme L_1, le saut était proportionnel à $sgn\ (u_i - v_{ji})$ et l'ajustement des centres était le même indépendamment de leur distance avec les échantillons. La QVAC avec la métrique $\rho(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{D} |u_i - v_i|^2$ a été étudiée en profondeur dans Pages et Printems (2003). Nous reprendrons donc le paramètre de saut

$$\gamma_n = \frac{4\ K^{3/D}}{4\ K^{3/D} + \pi^2\ n} \tag{4.26}$$

qui avait été déterminé par ces auteurs. Nous noterons par QVAC_1 et QVAC_2 les algorithmes de quantification vectorielle par apprentissage compétitif avec les normes L_1 et L_2 respectivement. Pareillement, les algorithmes de Lloyd seront notés Lloyd_1 et Lloyd_2.

Ayant trouvé une forme convenable pour le paramètre de saut, nous sommes maintenant en mesure de générer des scénarios à partir des méthodes de Lloyd et QVAC lorsque la

distribution est connue. Les résultats seront présentés au prochain chapitre. Pour l'instant, on se concentre sur le développement d'un algorithme qui minimise la distance de Wasserstein lorsque la distribution du vecteur aléatoire est inconnue.

4.3 Distribution inconnue

On supposera dans cette section que l'on possède un ensemble de données historiques à partir desquelles il n'est pas possible d'estimer la distribution du vecteur aléatoire avec confiance. Autrement dit, les données sont en quantité insuffisante ou ne semblent provenir d'aucune loi probabiliste évidente. Le nombre de données disponibles pour la quantification est limité, contrairement à la discrétisation d'une distribution aléatoire connue où il est possible de générer une quantité d'échantillons arbitrairement grande. Dans ce cas, il est préférable d'utiliser une méthode de partitionnement directement à partir de l'ensemble de données historiques.

Algorithme de Lloyd

Il existe deux versions de l'algorithme de Lloyd. La première correspond à celle présentée au chapitre précédent qui peut être utilisée pour minimiser la distance de Wasserstein lorsque la distribution U est continue et connue. La deuxième (que nous appellerons algorithme de Lloyd «discret» lorsqu'il y a matière à confusion) est plus couramment utilisée et permet de résoudre certains problèmes de partitionnement à partir d'un ensemble de données, comme ceux des k-moyennes ou des k-médianes. Elle représente la méthode de base pour ces problèmes et est parfois même appelée algorithme des k-moyennes (ou k-médianes). Nous présentons ici la variante de l'algorithme pour le problème des k-médianes, où l'on suppose que l'on veut générer K scénarios $\mathbf{v}_j$, $j = 1, ..., K$ à partir d'un ensemble de N données $\mathbf{u}_n$, $n = 1, ..., N$.

Algorithme de Lloyd (discret)

1. (*Initialisation*) On choisit un paramètre d'arrêt $\epsilon > 0$ arbitrairement petit. On tire K valeurs $\mathbf{u}_j$ aléatoirement et sans remises à partir de l'ensemble de données. On pose $s = 0$ et $\mathbf{v}_j^0 = \mathbf{u}_j$, $j = 1, ..., K$.
2. (*Sous-ensembles*) Pour chaque centre $\mathbf{v}_j^s$, on trouve parmi les données les sous-ensembles de points

$$\Omega_j^s = \{\mathbf{u}_n \mid \rho(\mathbf{u}_n, \mathbf{v}_j^s) \leq \min_l \rho(\mathbf{u}_n, \mathbf{v}_l^s) \ , \ i = 1, ...N, \ l = 1, ..., K\} \ , \tag{4.27}$$

où $\rho(\mathbf{u}_n, \mathbf{v}_j^s) = \sum_{i=1}^{D} |u_{ni} - v_{ji}|$.

3. (*Médianes*) On calcule la médiane arithmétique $\boldsymbol{\nu}_j^{1/2}$ de chaque sous-ensemble Ω_j^s et on pose $\mathbf{v}_j^{s+1} = \boldsymbol{\nu}_j^{1/2}$.
4. (*Critère d'arrêt*) Si $\rho(\mathbf{v}_j^{s+1},\ \mathbf{v}_j^s) < \epsilon\ \forall\ j$, alors on prend $\mathbf{v}_j^* = \mathbf{v}_j^{s+1}$. Sinon, on incrémente s de 1 unité ($s \leftarrow s+1$) et on retourne à la 2^e étape.

On entend par médiane arithmétique d'un ensemble Ω_j le point formé à partir de toutes les médianes marginales. Plus spécifiquement, si Ω_j possède N_j points, alors sa médiane $\boldsymbol{\nu}_j^{1/2} = (\nu_{j1}^{1/2}, ..., \nu_{jD}^{1/2})$ est définie par

$$\nu_{ji}^{1/2} = \begin{cases} 1/2\ (u_i^{(N_j/2)} + u_i^{(N_j/2+1)}) & \text{si } N_j \text{ est pair} \\ u_i^{(N_j+1)/2} & \text{si } N_j \text{ est impair} \end{cases} \tag{4.28}$$

pour $i = 1, ..., D$, et où $u_i^{(k)}$ correspond à la k^e plus petite valeur parmi tous les points pour la dimension i (i.e. la statistique d'ordre de rang k de la dimension i). Pour le problème des k-moyennes, l'algorithme de Lloyd est sensiblement le même, à l'exception de la métrique qui devient $\rho(\mathbf{u}_n, \mathbf{v}_j^s) = \sum_{i=1}^{D} |u_{ni} - v_{ji}|^2$ et des centres (calculés en 3^e étape) qui correspondent aux moyennes plutôt qu'aux médianes. Comme on l'avait fait dans le cas continu, on notera Lloyd_1 et Lloyd_2 les algorithmes avec les normes L_1 et L_2 respectivement. Les cas continus et discrets seront traités séparément, il ne devrait donc pas y avoir de confusion dans la notation.

On pourrait démontrer que les algorithmes de Lloyd génèrent une suite de solutions non croissantes, mais garantissent seulement la convergence vers un minimum local du problème des k-médianes (même pour le cas unidimensionnel). De plus, un exemple très simple permet de constater que ce minimum local peut se situer arbitrairement loin du minimum global. Considérons un vecteur aléatoire à deux dimensions $\mathbf{U}$ pouvant prendre les 4 valeurs suivantes avec probabilités égales :

$$\mathbf{U} = \begin{cases} \mathbf{u}_1 = (0,0) \\ \mathbf{u}_2 = (0,b) \\ \mathbf{u}_3 = (a,b) \\ \mathbf{u}_4 = (a,0) \end{cases} \qquad \text{avec probabilité } 1/4,$$

où a>b. Ces points forment un rectangle dans le plan représenté à la figure 4.6. Supposons que l'on cherche à approximer $\mathbf{U}$ par 2 scénarios qui minimisent la distance L_1. Selon l'initiali-

sation (aléatoire) de l'algorithme de Lloyd$_1$, on pourrait soit obtenir une solution représentée par les $*$ (minimum global) ou une autre représentée par les x (minimum local) [1].

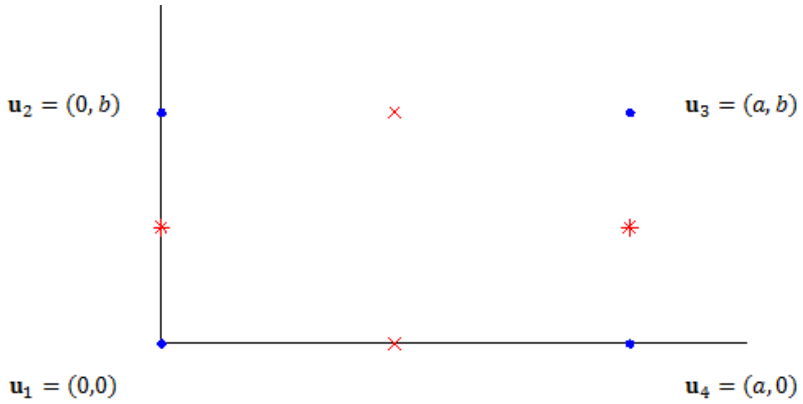

Figure 4.6 Scénarios représentant les solutions minimales locales (x) et optimales ($*$)

Les scénarios optimaux sont

$$\begin{aligned}\mathbf{v}_1^* &= 1/2\ (\mathbf{u}_1 + \mathbf{u}_2) = (0, b/2)\\ \mathbf{v}_2^* &= 1/2\ (\mathbf{u}_3 + \mathbf{u}_4) = (a, b/2)\ .\end{aligned}$$

Les ensembles définis par (3.38) pour ces scénarios sont alors

$$\begin{aligned}\Omega_1^* &= \{\mathbf{u}_1, \mathbf{u}_2\}\\ \Omega_2^* &= \{\mathbf{u}_3, \mathbf{u}_4\}\end{aligned}$$

et la distance de Wasserstein correspondante est

$$\begin{aligned}d_1^*(\mathbf{U},\mathbf{V}) &= 1/4\ \rho(\mathbf{u}_1, \mathbf{v}_1) + 1/4\ \rho(\mathbf{u}_2, \mathbf{v}_1) + 1/4\ \rho(\mathbf{u}_3, \mathbf{v}_2) + 1/4\ \rho(\mathbf{u}_4, \mathbf{v}_2)\\ &= 1/4\ |b - b/2| + 1/4\ |0 - b/2| + 1/4\ |b - b/2| + 1/4\ |0 - b/2|\\ &= b/2\ .\end{aligned}$$

D'autre part, les scénarios

$$\mathbf{v}_1^{\mathrm{x}} = 1/2\ (\mathbf{u}_1 + \mathbf{u}_3)$$

1. En fait, d'autres choix de minima locaux et optimaux existent. Par exemple, n'importe quel point sur les segments de droite reliant u_2 et u_3 ou reliant u_1 et u_2 auraient pu servir de minimum local.

$$\mathbf{v}_2^{\mathrm{x}} = 1/2\ (\mathbf{u}_2 + \mathbf{u}_4)$$

correspondent à des minima locaux et donnent

$$d_1^{\mathrm{x}}(\mathbf{U},\mathbf{V}) = a/2\ .$$

Or, le même résultat pourrait être obtenu avec un paramètre a arbitrairement supérieur à b. Par conséquent, le minimum local $d_1^{\mathrm{x}}(\mathbf{U},\mathbf{V})$ peut être arbitrairement supérieur au minimum global $d_1^*(\mathbf{U},\mathbf{V})$. Malgré tout, l'algorithme de Lloyd reste une méthode possible pour minimiser la distance de Wasserstein qui sera envisagée. Il peut engendrer des solutions plus ou moins bonnes dépendamment des instances et de l'initialisation des centres. Pour plus de robustesse, nous chercherons donc une méthode alternative qui permet de résoudre le problème des k-médianes, en espérant qu'elle peut combler certaines lacunes de l'algorithme de Lloyd.

Algorithme d'échange des centres

Une difficulté du problème des k-médianes provient du fait que le temps de convergence des algorithmes croît généralement rapidement en fonction du nombre de centres K et de données N. Une option possible est alors de considérer le problème approximatif des k-médoïdes.

Définition 4.3. *Soit* $\mathbf{U}$ *un ensemble de* N *données* $\mathbf{u}_n \in \mathbb{R}^D,\ n = 1, ..., N$, *et une métrique* $\rho(\mathbf{u},\mathbf{v})$. *Le* **problème des k-médoïdes** *consiste à chercher les centres* $\mathbf{v}_j \in \mathbf{U},\ j = 1, ..., K$, *appelés* **médoïdes**, *qui minimisent*

$$z(\mathbf{v}) = \sum_{n=1}^{N} \min_j\ \rho(\mathbf{u}_n, \mathbf{v}_j)\ . \tag{4.29}$$

La différence entre ce problème et celui des k-médianes est que les centres $\mathbf{v}_j$ appartiennent à $\mathbf{U}$ et la métrique utilisée est arbitraire. Cependant, nous considérerons dans cette section le problème des k-médoïdes avec la métrique $\rho(\mathbf{u},\mathbf{v}) = \sum_{i=1}^{n} |u_i - v_i|$. Dans ce cas, les deux problèmes sont identiques à l'exception de la contrainte supplémentaire exigeant que les centres correspondent à des données pour les k-médoïdes. L'avantage principal de cette contrainte est qu'elle nous assure de trouver des scénarios entiers lorsque l'on discrétise un vecteur aléatoire ne prenant que des valeurs entières, contrairement à l'algorithme de Lloyd qui trouve des scénarios $\mathbf{v}_j \in \mathbb{R}^D$, peu importe la distribution. De plus, le nombre de partitions que l'on peut former correspond au nombre de combinaisons possibles de K valeurs parmi N, qui est considérablement inférieure au nombre de Stirling, i.e. $\binom{N}{K} \ll S(N, K)$. Par conséquent, les algorithmes des k-médoïdes convergent souvent plus rapidement que ceux des

k-médianes (par exemple, voir Velmurugan et Santhanam (2010)).

Nous résoudrons dans cette section le problème des k-médoïdes à l'aide d'une méthode basée sur l'algorithme d'échange des centres de Korupolu *et al.* (2000). On notera par $\mathbb{V} = \{\mathbf{v}_1, ..., \mathbf{v}_K\}$ l'ensemble des centres.

Algorithme d'échange des centres

1. (*Initialisation*) On tire K données $\mathbf{u}_j,\ j = 1, ..., K$, aléatoirement et sans remises parmi l'ensemble $\mathbf{U}$. On pose $s = 0$ et $\mathbf{v}_j^0 = \mathbf{u}_j\ \forall\ j$, puis on calcule la matrice des distances

$$D = \begin{pmatrix} \rho(\mathbf{u}_1, \mathbf{u}_1) & \rho(\mathbf{u}_1, \mathbf{u}_2) & \cdots & \rho(\mathbf{u}_1, \mathbf{u}_N) \\ \rho(\mathbf{u}_2, \mathbf{u}_1) & \rho(\mathbf{u}_2, \mathbf{u}_2) & \cdots & \rho(\mathbf{u}_2, \mathbf{u}_N) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(\mathbf{u}_N, \mathbf{u}_1) & \rho(\mathbf{u}_N, \mathbf{u}_2) & \cdots & \rho(\mathbf{u}_N, \mathbf{u}_N) \end{pmatrix}$$

2. (*Solution actuelle*) On évalue le coût $z(\mathbf{v}^s)$ de la solution actuelle (à l'aide de (4.1)) et on pose $z_{min}^s = z(\mathbf{v}^s)$.
3. (*Échange des centres*)

   pour $j = 1, ..., K$
   pour $n = 1, ..., N$
   si $\mathbf{u}_n \notin \mathbb{V}$ **alors**
   (i) On échange temporairement $\mathbf{u}_n$ et $\mathbf{v}_j^s$, i.e. $\mathbf{u}_n$ devient un centre tandis $\mathbf{v}_j^s$ n'en est plus un.
   (ii) On évalue le coût z_{jn}^s de la solution obtenue par l'échange en (i).
   (iii) Si $z_{jn}^s < z_{min}^s$, alors $z_{min}^s \leftarrow z_{jn}^s$, $j_{min} \leftarrow j$ et $n_{min} \leftarrow n$.
   (iv) $\mathbf{v}_j^s$ et $\mathbf{u}_n$ redeviennent respectivement un centre et un point normal.
   fin si
   fin pour
   fin pour

4. (*Terminaison*) Si $z_{min}^s = z_{min}^{s-1}$, alors $\mathbf{v}^* = \mathbf{v}^s$. Sinon, on effectue définitivement l'échange entre $\mathbf{u}_{i_{min}}$ et $\mathbf{v}_{j_{min}}^s$, $s \leftarrow s + 1$ et on retourne à l'étape 2.

Pour le problème des k-médianes, la matrice des distances est énorme étant donné le grand nombre de centres possibles. La distance entre les médianes et les données est donc normalement calculée itérativement. En dimension élevée, le temps de convergence peut devenir prohibitif puisque le calcul des métriques devient plus long. Un des avantages du problème

des k-médoïdes est la possibilité de ne calculer la matrice des distances qu'une seule fois dès l'initialisation de l'algorithme, ce qui nous permet d'économiser beaucoup de temps à l'étape (ii). Le temps de convergence total est $O(N^2K^2)$ et dépend peu de la dimension du problème.

Le choix du nombre de centres est une décision cruciale pour le problème des k-médoïdes. D'une part, un nombre plus élevé de centres conduira vers une solution inférieure aux dépens d'une convergence plus lente de l'algorithme utilisé. D'autre part, les centres correspondent aux scénarios dont la quantité affecte également le temps de résolution du programme stochastique. Le théorème suivant représente un outil qui peut nous aider à prendre une décision plus éclairée par rapport au nombre de centres choisi.

Théorème 4.1. *L'algorithme d'échange des centres résout le problème des k-médoïdes avec un facteur d'approximation* $(\alpha, \beta) = (1 + 5/\epsilon, 3 + \epsilon)$. *C'est-à-dire qu'il utilise au plus* βK *centres pour trouver un partitionnement qui est inférieur ou égal à* α *fois la solution optimale du problème des k-médoïdes avec* K *centres (et où* $N \geq \beta K$*).*

Preuve Voir Korupolu *et al.* (2000).

Supposons par exemple que l'on hésite entre résoudre le PS à l'aide de 50 ou 500 scénarios. Il y a un facteur $\beta = 10$ entre ces quantités de scénarios. On a donc $\epsilon = 7$ et le théorème 4.1 nous indique alors que le coût de la solution avec 500 scénarios sera au plus $(1 + 5/7)$ fois supérieure à la solution optimale du problème avec 50 centres. Même si ce résultat ne semble pas très convaincant à première vue, il faut se rappeler que la solution optimale est rarement atteinte pour les problèmes des k-médoïdes. De plus, la borne supérieure α est un résultat théorique qui correspond à la pire des situations et ne sera probablement jamais atteint en réalité.

Il existe une énorme quantité d'algorithmes possédant chacun leurs forces et leurs faiblesses qui ont été développés pour résoudre les problèmes des k-médianes ou k-médoïdes. En général, un temps de convergence plus lent sera compensé par une solution optimale moyenne de meilleure qualité. Si le PS nécessite peu de scénarios, il serait alors possible de considérer une méthode plus précise comme celle de Likas *et al.* (2002). Au contraire, s'il faut générer plusieurs scénarios à partir d'une grande quantité de données, des algorithmes de partitionnement plus rapides comme celui de Indyk (1999), qui possède un temps de convergence linéaire en N, devraient être considérés. Lorsque N est trop grand, il est aussi possible de faire appel à des processus préliminaires pour réduire la quantité de données considérée par l'algorithme de partitionnement (par exemple, voir Chen (2009)).

Nous avons maintenant en main divers algorithmes qui nous permettent de générer des scénarios dans les cas où la distribution continue est connue et dans ceux où elle ne l'est pas. La seule situation qui n'a pas encore été envisagée est celle des distributions discrètes. Pour générer des scénarios dans un tel cas, il y a deux options possibles basées sur les méthodes étudiées dans ce mémoire. La première est d'utiliser l'algorithme d'échange des centres directement à partir des données. L'avantage de cette procédure est qu'elle traite le problème des k-médoïdes : les scénarios correspondront donc nécessairement à des réalisations du vecteur aléatoire discret. Si la distribution peut être estimée, la seconde option consiste à considérer la relaxation linéaire du vecteur aléatoire discret et d'utiliser la QVAC (ou l'algorithme de Lloyd continu en 1 dimension) pour générer les scénarios. Dans ce cas, les scénarios ne correspondront pas à des réalisations du vecteur aléatoire, mais pourraient tout de même être utilisés pour résoudre le PS.

CHAPITRE 5

RÉSULTATS ET ANALYSE

5.1 Distribution connue

On considère d'abord les problèmes pour lesquels il est possible d'estimer la distribution du vecteur aléatoire continu avec confiance. Par souci de concision, on supposera tout simplement que la distribution de $\mathbf{U}$ est connue. On utilisera l'algorithme de Lloyd pour les variables aléatoires en une dimension et les méthodes de quantification vectorielle par apprentissage compétitif présentées au chapitre précédent pour le cas multidimensionnel. La QVAC avec la norme L_1 (i.e. la QVAC_1) sera utilisée avec le facteur de saut trouvé à partir de la distribution normale

$$\gamma_n = 0,8\ \sigma_i \times \frac{175}{175+n} \qquad \text{pour } D = 2,\ 3, ... \tag{5.1}$$

tandis qu'on aura

$$\gamma_n = \frac{4\ K^{3/D}}{4\ K^{3/D} + \pi^2\ n} \qquad \text{pour } D = 2,\ 3, ... \tag{5.2}$$

pour la QVAC_2, avec $n = 1, ...N$. Sauf avis contraire, les implémentations de la QVAC sont toutes effectuées avec $N = 3 \times 10^6$ itérations.

5.1.1 Effet des instances

On a commencé par tester les algorithmes pour les distributions continues avec la norme L_1, c'est-à-dire ceux de Lloyd_1 (continu) et QVAC_1, sur une loi multinormale dont les distributions marginales sont non corrélées et suivent toutes une distribution $\mathcal{N}(0,1)$. Les résultats sont présentés dans le tableau 5.1 pour différentes instances, variant selon la dimension et le nombre de centres. Lorsque $D = 1$, l'algorithme de Lloyd_1 est utilisé tandis que les résultats pour $D = 2$ et 10 proviennent de la QVAC_1. Les valeurs de $d_1(\mathbf{U}, \mathbf{V})$ et du temps de convergence correspondent à une moyenne des résultats sur 10 implémentations des algorithmes. La signification de $\hat{\mu}_{E[\text{V}]}$ et $\hat{\sigma}^2_{E[\text{V}]}$ sera expliquée sous peu. Nous analysons plus en détail certains aspects de nos méthodes de génération de scénarios dans ce qui suit.

Temps de convergence

L'algorithme de Lloyd_1 a été implémenté sur MATLAB R2013b, tandis que la QVAC a été exécutée à partir de JAVA. Le temps de convergence pour la méthode de Lloyd_1 a donc

Tableau 5.1 Comparaison des résultats de l'algorithme de Lloyd_1 (pour $D = 1$), de la QVAC_1 (pour $D = 2$ et 10) et de l'échantillonnage pur (EP).

		Lloyd_1 / QVAC_1				EP
D	K	$d_1(\mathbf{U}, \mathbf{V})$	temps (s)	$\hat{\mu}_{E[\text{V}]}$	$\hat{\sigma}^2_{E[\text{V}]}$	$d_1(\mathbf{U}, \mathbf{V})$
1	5	0,218	-	$-9,67 \times 10^{-17}$	$2,39 \times 10^{-32}$	0,338
1	50	0,025	-	-	-	0,056
1	100	0,013	-	-	-	0,030
2	5	0,860	2,33	$3,42 \times 10^{-3}$	$1,34 \times 10^{-4}$	1,153
2	50	0,299	8,56	-	-	0,405
2	100	0,217	16,02	-	-	0,286
10	50	5,736	65,08	$-2,33 \times 10^{-3}$	$4,03 \times 10^{-5}$	6,355
10	100	5,394	99,89	-	-	5,844
10	500	4,722	376,20	-	-	4,902

été omis du tableau puisqu'il ne peut pas réellement être comparé à celui de la QVAC_1. Les tests ont tous été implémentés sur un ordinateur personnel muni d'un processeur Intel Core i5 de 2,50 GHz et mémoire vive de 8,00 GB. Les temps de convergence sont seulement indiqués à titre de comparaison entre les méthodes et pourraient facilement être améliorés par un système plus performant. On constate que le temps pour la QVAC_1 est environ proportionnel à $K \cdot D$, ce qui semble indiqué un temps de convergence $O(KD)$. Pour sa part, l'échantillonnage pur ne consiste qu'à échantillonner quelques valeurs de la distribution et se fait quasi instantanément. Nous avons utilisé la QVAC_1 avec un nombre fixe d'itérations, mais la convergence pourrait être accélérée en fixant un critère d'arrêt lorsque la différence entre les solutions successives devient trop basse.

Espérance des scénarios

Il est également intéressant d'observer la capacité de la méthode de QVAC_1 à estimer l'espérance de la distribution correctement. Pour ce faire, on introduit les estimateurs suivants :

$$\hat{\mu}_{E[\text{V}]} = \frac{1}{T} \sum_{t=1}^{T} E[\text{V}_{1t}] \tag{5.3}$$

$$\hat{\sigma}_{E[\text{V}]} = \frac{1}{T-1} \sum_{t=1}^{T} \left(E[\text{V}_{1t}^2] - E[\text{U}]^2 \right) = \frac{1}{T-1} \sum_{t=1}^{T} \left(E[\text{V}_{1t}]^2 \right) \ , \tag{5.4}$$

où T correspond au nombre d'implémentations observées, les V_{1t} représentent les valeurs de

la 1^e dimension des scénarios $\mathbf{V}_t$ et

$$E[\mathrm{V}_1] = \sum_{j=1}^{K} p_j \cdot v_{j1} \ . \tag{5.5}$$

Autrement dit, $E[\mathrm{V}_{1t}]$ est la moyenne de la 1^{re} distribution marginale des scénarios obtenus à l'implémentation t. Quelques résultats de ces estimateurs sont présentés dans le tableau 5.1 sur un ensemble de $T = 40$ implémentations de l'algorithme. On remarque que les valeurs de $\hat{\mu}_{E[\mathrm{V}]}$ sont très près de 0 (ce qui correspond à l'espérance des distributions marginales $\mathcal{N}(0,1)$) et varient peu autour de la moyenne. Les résultats étant déjà excellents pour les plus petites valeurs de K, nous n'avons pas jugé nécessaire de calculer les estimateurs lorsqu'il y a plus de scénarios, puisque l'on s'attend à ce que l'estimation de l'espérance soit encore meilleure dans ces cas. On en déduit donc que les méthodes Lloyd_1 et QVAC_1 génèrent des scénarios dont l'espérance correspond presque exactement à celle de la distribution normale, même s'il ne s'agit pas de l'objectif principal de la méthode.

Dimension

L'échantillonnage pur a été défini à la section 3.2 et consiste simplement à tirer les scénarios au hasard parmi la distribution estimée. Les probabilités sont ensuite déterminées par la proportion d'échantillons sur un grand nombre de tirages dont la distance avec chaque scénario est la plus courte. Puisque cette méthode est beaucoup plus simple et rapide que celles présentées dans ce mémoire, il est nécessaire de comparer nos résultats avec ceux de l'échantillonnage pur pour déterminer si l'effort en vaut la peine.

En une dimension, l'algorithme de Lloyd_1 donne des résultats arbitrairement près du minimum global. Nous avons utilisé l'algorithme avec 2000 itérations, on peut donc s'attendre à ce que les valeurs de $d_1(\mathbf{U},\mathbf{V})$ obtenues correspondent approximativement à la solution optimale. Pour le cas multidimensionnel, on remarque effectivement à partir du tableau 5.1 que la distance de Wasserstein obtenue par la QVAC_1 est toujours inférieure à celle de l'échantillonnage pur. La relation entre les deux méthodes et la dimension est représentée sur la figure 5.1 pour des distributions $\mathcal{N}(0,1)$ non corrélées discrétisées par 50 scénarios. On observe que la distance de Wasserstein augmente exponentiellement en fonction de la dimension tant pour la QVAC_1 que pour l'échantillonnage pur. Bien que l'écart absolu augmente légèrement, le gain relatif ($G_{\%}$) de la quantification vectorielle par rapport à l'échantillonnage pur, que nous

définissons par

$$G_{\%} = \frac{d_1^{EP}(\mathbf{U},\mathbf{V}) - d_1^{QVAC}(\mathbf{U},\mathbf{V})}{d_1^{EP}(\mathbf{U},\mathbf{V})} \times 100 \tag{5.6}$$

diminue avec la dimension (d_1^{EP} et d_1^{QVAC} sont les distances de Wasserstein obtenues par échantillonnage pur et avec la QVAC respectivement). Pour $D \geq 4$, le gain se situe autour de 10% et décroît lentement.

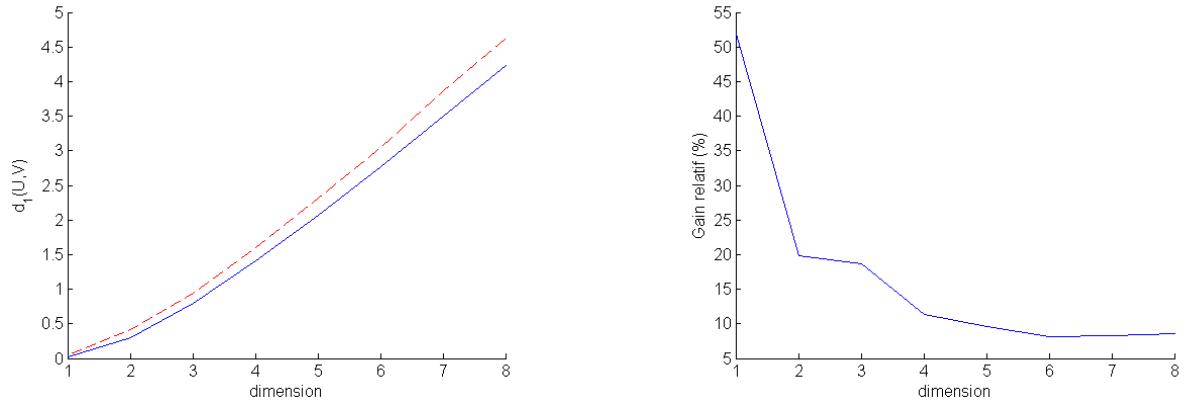

Figure 5.1 Distance de Wasserstein pour les méthodes de Lloyd$_1$/QVAC$_1$ (ligne pleine) et d'échantillonnage pur (ligne pointillée) en fonction de la dimension (figure de gauche). Gain relatif en fonction de la dimension (figure de droite).

Nombre de scénarios

La distance de Wasserstein diminue évidemment en fonction du nombre de scénarios, comme on peut le constater sur la figure 5.2 qui représente les valeurs de $d_1(\mathbf{U},\mathbf{V})$ pour $D = 3$ distributions marginales $\mathcal{N}(0,1)$ non corrélées en fonction du nombre de scénarios. Le gain relatif est représenté à la figure 5.2 et se situe autour 13,5 % indépendamment de K. De plus, pour une dimension donnée, on déduit la relation suivante à partir du tableau 5.1 :

$$K^{1/D} \cdot d_1(\mathbf{U},\mathbf{V}) \approx cste \ , \tag{5.7}$$

c'est-à-dire qu'une augmentation du nombre de scénarios d'un facteur α réduit la distance de Wasserstein d'environ $\alpha^{1/D}$. Ainsi, plus la dimension est élevée, plus il devient difficile de minimiser la distance de Wasserstein. Ce propos est également confirmé par la figure 5.1. Nous chercherons donc des façons de réduire la dimension effective de notre problème un peu plus tard dans l'espoir d'améliorer nos méthodes de génération de scénarios en dimension élevée.

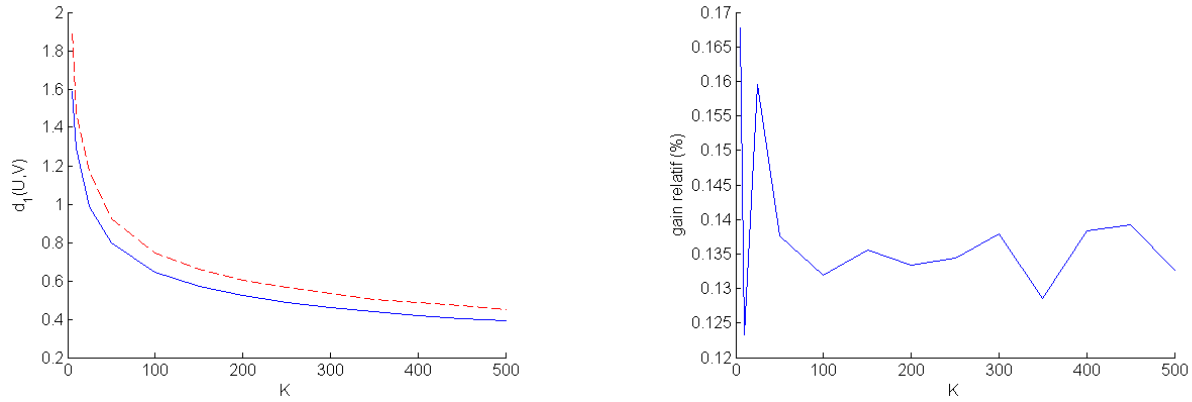

Figure 5.2 Distance de Wasserstein pour la QVAC_1 (ligne pleine) et d'échantillonnage pur (ligne pointillée) en fonction du nombre de médianes (figure de gauche). Gain relatif en fonction du nombre de médianes (figure de droite).

Corrélations

Jusqu'à présent, nous avons supposé que les distributions marginales étaient non corrélées. Nous allons maintenant analyser l'effet des corrélations sur les méthodes de génération de scénarios. Pour utiliser la QVAC_1, il est nécessaire de connaître ou d'avoir estimé la distribution du vecteur aléatoire de notre problème. La première conséquence d'une dépendance substantielle entre les distributions marginales est que l'on doit évaluer la matrice des corrélations, ce qui représente $D(D-1)/2$ paramètres supplémentaires. Nous supposerons encore une fois que les corrélations ont été estimées *a priori*. Leur influence sur la QVAC_1 et l'échantillonnage pur a été testée sur un vecteur de 3 distributions $\mathcal{N}(0,1)$ quantifiées par 50 scénarios dont les coefficients de corrélations linéaires ($\rho_{ij},\ i \neq j$) sont tous égaux et varient entre 0 et 1 (voir figure 5.3). On observe que $d_1(\mathbf{U},\mathbf{V})$ décroît en fonction des corrélations, surtout lorsque celles-ci sont fortes ($> 0,7$) et qu'elles font croître le gain relatif de la QVAC_1 par rapport à l'échantillonnage pur (voir figure 5.3). Lorsque les distributions marginales sont parfaitement corrélées, les réalisations s'alignent le long d'une droite et notre problème devient unidimensionnel. Les corrélations diminuent donc la dimension effective de notre problème et permettent d'observer des gains plus importants de la QVAC.

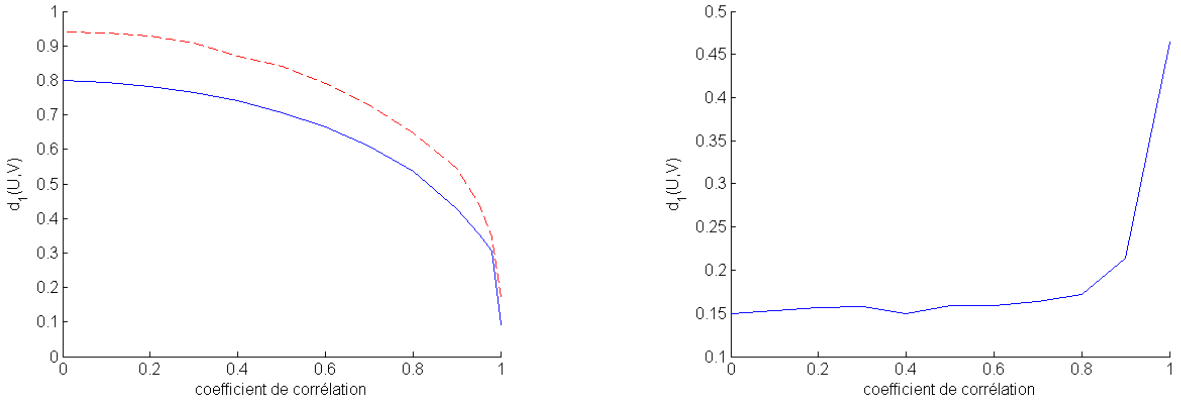

Figure 5.3 Distance de Wasserstein pour les méthodes de QVAC_1 (ligne pleine) et d'échantillonnage pur (ligne pointillée) en fonction des coefficients de corrélations linéaires (figure de gauche). Gain relatif en fonction des corrélations (figure de droite).

Autres distributions

Le paramètre de la QVAC_1 donné par (5.1) a été obtenu à partir d'un vecteur aléatoire dont les distributions marginales non corrélées suivaient toutes une loi normale. Or, nous utiliserons aussi ce paramètre de saut pour discrétiser d'autres distributions dans la suite de ce chapitre. Nous avons donc vérifié qu'il menait également à de bonnes solutions pour le problème de minimisation de la distance de Wasserstein pour quelques autres lois probabilistes. Pour ce faire, nous avons généré 50 scénarios à l'aide de la QVAC_1 sur des vecteurs aléatoires à 3 dimensions non corrélées dont les distributions marginales ont toutes été choisies de manière à posséder une variance égale à 1. Les résultats sont présentés dans le tableau 5.2.

Tableau 5.2 Distances de Wasserstein obtenues avec la QVAC_1 et l'échantillonnage pur (EP) pour différentes distributions.

Loi multinomiale	Distribution marginale	QVAC_1	EP	Gain relatif (%)
Multinormale	$\mathcal{N}(0,1)$	0,797	0,939	15,08
Uniforme	$\mathcal{U}(0,\sqrt{12})$	0,686	0,833	17,65
Gamma	$\mathcal{G}(4,\frac{1}{2})$	0,744	0,892	16,49
Exponentielle	$Exp(1)$	0,615	0,762	19,23
Log-normale	Log-$\mathcal{N}(0;\ 0,6937)$	0,614	0,768	20,00

Bien que le paramètre de saut ait été ajusté à partir d'une loi multinormale, les gains sur la distance de Wasserstein obtenus par rapport à l'échantillonnage pur se situent entre 15 et

20 % pour toutes les distributions. On peut donc se satisfaire de la forme de γ_n donnée par (5.1) afin de minimiser la distance de Wasserstein pour les prochaines distributions considérées dans ce chapitre.

Résumons brièvement les conclusions apportées par cette sous-section. Nous avons comparé les distances de Wasserstein obtenues par l'échantillonnage pur avec celles de l'algorithme de Lloyd_1 (continu) en une dimension et la QVAC_1 dans le cas multidimensionnel en fonction de 5 aspects. Mis à part le temps de convergence plus rapide de l'échantillonnage pur, toutes les autres caractéristiques des instances favorisent l'utilisation des algorithmes de Lloyd_1 et QVAC_1. Ces derniers offrent de meilleurs résultats que l'échantillonnage pur indépendamment de la dimension, du nombre de scénarios, etc. Le gain relatif diminue en fonction de la dimension tandis qu'il est environ constant indépendamment du nombre de scénarios et augmente en fonction des corrélations.

Nous n'avons pas encore testé nos algorithmes avec la distance euclidienne, mais on peut se convaincre que le comportement des algorithmes Lloyd_2 et QVAC_2 en fonction des caractéristiques étudiées ci-haut sera similaire à celui avec la norme L_1. En général, la différence entre les solutions de deux implémentations est certainement moins grande en utilisant deux normes différentes qu'en se servant de deux algorithmes distincts. Nos méthodes seront testées avec la norme L_2 prochainement.

5.1.2 Méthodes de réduction de la dimension

Nous avons vu qu'il devenait difficile de minimiser la distance de Wasserstein en dimension élevée (i.e. pour $D \geq 4$) comparativement au cas unidimensionnel. Il est donc important de réfléchir à des méthodes qui permettraient de réduire $d_1(\mathbf{U}, \mathbf{V})$ lorsque D devient grand. Nous cherchons donc dans cette sous-section des manières de réduire la dimension effective de notre problème.

Heuristique de quadrillage

Pour tirer profit du fait qu'il est plus facile de discrétiser une variable aléatoire en 1 dimension, une heuristique possible consiste à générer les scénarios de chaque distribution marginale à partir de l'algorithme de Lloyd_1 et de les combiner pour obtenir des scénarios multidimensionnels. Cependant, il s'agit effectivement d'une heuristique, puisque les centres sont alors placés sous forme d'un quadrillage hyperdimensionnel, ce qui ne correspond généralement pas à la configuration optimale. On peut apercevoir sur la figure 5.4 l'emplacement des 49 scéna-

rios obtenus par la QVAC_1 pour une loi binormale avec vecteur des moyennes $\boldsymbol{\mu} = (0,0)^T$ et matrice des covariances $\Sigma = I_2$ en comparaison avec 7 scénarios d'une $\mathcal{N}(0,1)$ placés en grille.

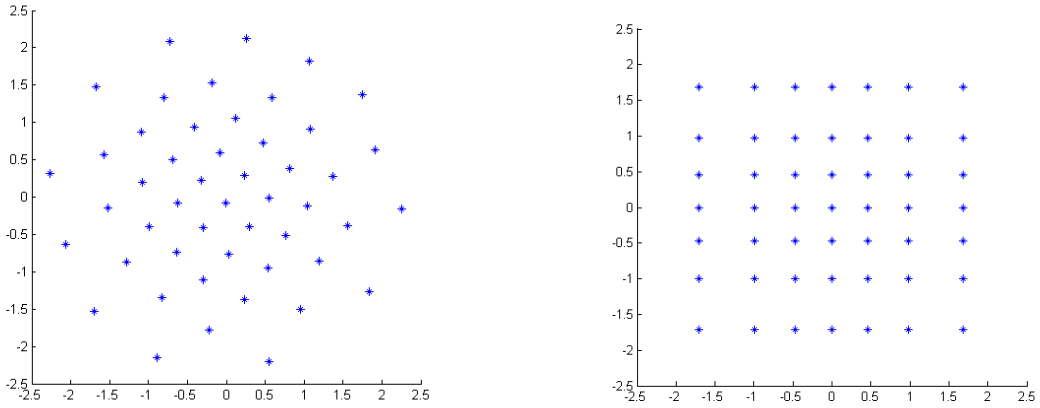

Figure 5.4 Représentation de 49 scénarios discrétisant un loi multinormale standard bidimensionnelle non corrélée obtenus par QVAC_1 et par l'heuristique de quadrillage.

On trouve que les distances de Wasserstein obtenues pour les méthodes de QVAC_1 et par l'heuristique de quadrillage pour 49 scénarios en 2 dimensions sont respectivement $d_1^{QVAC_1}(\mathbf{U},\mathbf{V}) = 0,3021$ et $d_1^{quad}(\mathbf{U},\mathbf{V}) = 0,3231$. Ainsi, même lorsque les distributions sont indépendantes, la solution optimale ne peut pas être obtenue en générant des scénarios pour chaque distribution marginale individuellement. L'heuristique génère toutefois des scénarios dont la distance de Wasserstein est relativement près de celle obtenue par la QVAC_1. Notons que l'on peut choisir un nombre de centres différent pour chaque dimension. Par exemple, il serait possible de discrétiser 3 distributions marginales par K_1, K_2 et K_3 centres pour obtenir un ensemble de $K = K_1 \cdot K_2 \cdot K_3$ scénarios. Même si elle peut être une bonne alternative à la QVAC_1 dans certains cas, cette méthode possède ses limites. Tout d'abord, elle succombe facilement au «fléau de la dimensionnalité». Supposons que l'on souhaite produire des scénarios pour un vecteur aléatoire en 20 dimensions. En ne générant alors que 2 centres par dimension, on se retrouverait avec un ensemble de $2^{20} \approx 10^6$ scénarios. Même en dimension moins élevée, la quantité de scénarios peut rapidement devenir prohibitive. Ensuite, l'heuristique de quadrillage ne permet pas de générer des scénarios pour un vecteur aléatoire en considérant les corrélations entre ses composantes. Si les distributions marginales ne sont pas indépendantes, alors la génération des scénarios pour chacune d'entre elles prise individuellement est loin d'être optimale. Notons pour terminer que même dans le cas de distributions uniformes non corrélées avec la norme L_1, les scénarios optimaux ne forment pas un quadrillage (voir figure 5.5). Le cas d'une distribution uniforme avec la distance euclidienne et

$K = n^D,\ n \in \mathbb{N}$ est le seul où la solution optimale engendre un quadrillage.

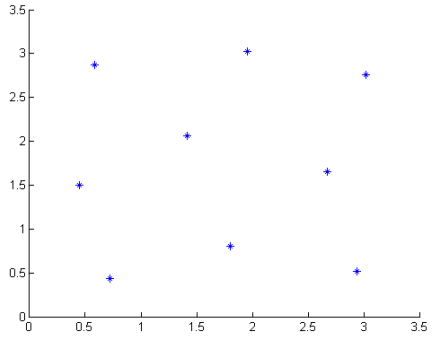

Figure 5.5 Représentation de 9 scénarios discrétisant des distributions $\mathcal{U}(0, \sqrt{12})$ non corrélées obtenus par QVAC_1

Copules

La génération de scénarios dans le cas multidimensionnel se résume principalement à quantifier les distributions marginales en tenant compte de leurs corrélations. L'heuristique de quadrillage offre la possibilité de générer des scénarios beaucoup plus rapidement que la QVAC_1, mais ne permet pas de réduire la distance de Wasserstein. En outre, elle ne permet pas de considérer les dépendances entre les variables aléatoires.

Les distributions marginales sont certainement mieux discrétisées lorsqu'elles sont considérées individuellement. On aimerait utiliser une méthode qui tire avantage de cette caractéristique tout en respectant la matrice des corrélations. On se servira donc des copules, qui permettent justement de discrétiser les distributions marginales et les corrélations séparément.

Définition 5.1. *Soit $F_{\mathbf{U}}$ la fonction de répartition de $\mathbf{U}$ dont les fonctions de répartitions marginales sont $F_1, ...,\ F_D$. Une* **copule** *est une fonction $C : [0,1]^D \to [0,1]$ telle que*

$$F_{\mathbf{U}}(u_1, u_2, ..., u_D) = C\left(F_1(u_1),\ F_2(u_2), ..., F_D(u_D)\right)\ . \tag{5.8}$$

Théorème 5.1. (Théorème de Sklar) *Il existe toujours une copule satisfaisant la relation* (5.8)*. De plus, si les fonctions de répartition $F_i,\ i = 1, ..., D$ sont toutes continues, alors la copule est unique.*

Preuve Voir Sklar (1996).

Ainsi, si les fonctions répartition sont continues, on peut poser $y_i = F_i(u_i)$ et on trouve

$$C(y_1, y_2, ..., y_D) = F_{\mathbf{U}}\left(F_1^{-1}(y_1), F_2^{-1}(y_2), ..., F_D^{-1}(y_D)\right) \ . \tag{5.9}$$

Le but des copules n'est pas de trouver la forme explicite de la fonction (5.9) (ce qui est généralement très difficile). Elles nous serviront plutôt à découpler les distributions marginales pour la génération de scénarios.

On suppose que l'on veut générer des scénarios à partir de la distribution **U**. Notons tout d'abord que si U_i est une variable aléatoire de fonction de répartition F_i, alors $\mathrm{Y}_i = F_i(\mathrm{U}_i)$ suit une loi $\mathcal{U}(0,1)$. Après l'application de F_i, il ne reste donc plus aucune information à propos de la distribution initiale U_i. Par conséquent, $C(y_1, y_2, ..., y_D)$ ne dépend que des corrélations entre les variables aléatoires, et non de leurs distributions marginales. Plutôt que de chercher à discrétiser $F_{\mathbf{U}}$ directement, les copules nous permettront de générer les scénarios en trois temps :

1. On quantifie les distributions marginales U_i
2. On discrétise $C(y_1, y_2, ..., y_D)$ et on associe les centres de chaque distribution
3. On trouve les probabilités associées à chaque scénario.

La première étape se fait simplement en utilisant la méthode de Lloyd_1. Les scénarios marginaux obtenus en première étape seront notés $v_{ji},\ j = 1, ..., K,\ i = 1, ..., D$ et leurs probabilités respectives p_{ji}.

La discrétisation de la copule permet d'approximer les corrélations entre les distributions marginales. On utilise le fait que C correspond à une fonction de répartition dont les distributions marginales suivent toutes des lois $\mathcal{U}(0,1)$. On a supposé dans cette section que la distribution du vecteur aléatoire est connue. Les observations de C correspondent donc à $(F_1(u_1),\ F_2(u_2), ..., F_D(u_D))$, où $(u_1, ..., u_D)$ est un échantillon de **U**. Par exemple, la figure 5.6 montre des échantillons de distributions $\mathrm{U}_1 \sim \mathcal{N}(0,1)$ avec covariance $\sigma_{12} = 0,7$. La figure 5.6 correspond au résultat après l'application des fonctions de répartition marginales, c'est-à-dire à des observations de la copule.

Puisque $Var\left[\mathcal{U}(0,1)\right] = 1/12$, on discrétisera les copules par la QVAC_1 avec le paramètre de

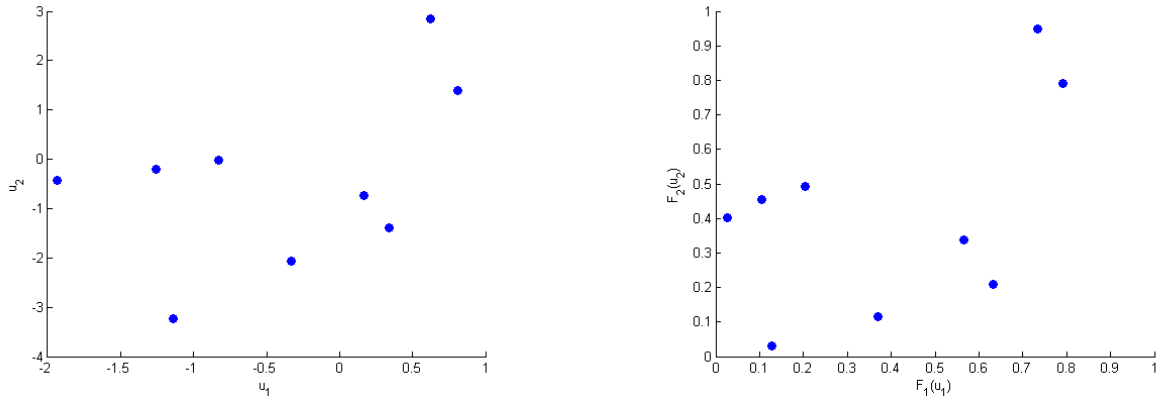

Figure 5.6 Échantillons de distributions multinormales (figure de gauche). Observations obtenues après application des fonctions de répartition marginales (figure de droite).

saut

$$\gamma_n = 0,8 \times \frac{1}{\sqrt{12}} \times \frac{175}{175+n} \ . \tag{5.10}$$

La quantification des copules sert à déterminer quels scénarios marginaux devraient être associés. Si $\mathbf{y}_j = (y_{j1}, ..., y_{jD})$ est un centre représentant la copule, alors un scénario sera donné par

$$\mathbf{v}_j = (v_{rang(y_{j1}),1},\ v_{rang(y_{j2}),2},\ \ ...\ , v_{rang(y_{jD}),D}) \ , \tag{5.11}$$

où $rang(y_{ji})$ correspond au rang de y_{ji}, que nous définissons formellement ci-dessous.

Définition 5.2. *Soit $y_{1i}, ..., y_{Ki}$ les valeurs des centres représentant la copule pour une dimension i donnée. Le* **rang** *de $y_{ji},\ j = 1, ..., K$, est défini par*

$$rang(y_{ji}) = \sum_{l=1}^{K} \mathbb{I}\{y_{li} \leq y_{ji}\} \ , \tag{5.12}$$

où $\mathbb{I}\{y_{li} \leq y_{ji}\} = 1$ si $y_{li} \leq y_{ji}$ et 0 sinon.

La figure 5.7 montre 5 centres de la copule pour la distribution $U_1 \sim \mathcal{N}(0,1)$ avec $\sigma_{12} = 0,7$ obtenus par la QVAC_1. Les lignes pointillées servent uniquement à faciliter la visualisation du rang de chaque centre. On observe par exemple que le point $\mathbf{y}_1 = (0,125; 0,318)$ possède des rangs $rang(y_{11}) = 1$ et $rang(y_{12}) = 1$. Un scénario sera donc donné par

$$\mathbf{v} = (v_{rang(y_{11}),1}, v_{rang(y_{12}),2}) = (v_{11}, v_{12}) \ .$$

Pour cet exemple, la discrétisation de la copule établit donc la règle d'associer les 2 plus petits scénarios marginaux ensemble, le 2^e avec le 4^e, le 3^e avec le 2^e, etc. L'association des

scénarios marginaux pour l'ensemble des points est représentée sur la figure 5.7.

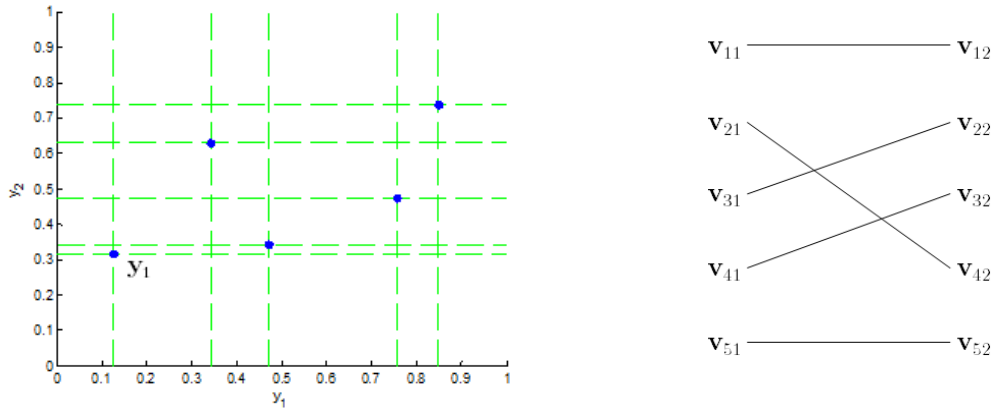

Figure 5.7 Centres des copules (figure de gauche). Association des copules en fonction du rang des centres (figure de droite).

Pour associer les probabilités aux scénarios, on n'a qu'à tirer plusieurs valeurs de la distribution et compter le nombre de fois où la médiane est plus près de l'échantillon. Cette procédure est analogue à la manière de calculer les probabilités pour l'échantillonnage pur. Plus d'information sur la génération de scénarios par les copules est disponible dans Kaut et Wallace (2011) ou Kaut (2013). Ces auteurs avaient obtenu des résultats encourageants lorsque les copules étaient utilisées conjointement avec la méthode de correspondance des moments.

Dans notre cas, on constate cependant à partir du tableau 5.3 que les résultats de la méthode des copules sont décevants, puisque la distance de Wasserstein est nettement supérieure à celle obtenue par la QVAC_1. Une partie de l'explication provient du fait que les distributions marginales trouvées en 1^e étape ne peuvent être conservées lorsque l'on jumelle les centres. En deux dimensions par exemple, on forme 5 scénarios à partir des 5 réalisations pour chacune des distributions marginales, plutôt que d'en former 25 en multipliant leur probabilité (comme pour l'heuristique de quadrillage), ce qui permettrait de conserver les lois de probabilité marginales.

Une deuxième cause possible des résultats médiocres de la méthode des copules est la suivante : les scénarios ont tendance à se localiser plus près de la médiane au fur et à mesure que la dimension augmente (voir Pages et Printems (2003)). C'est-à-dire que pour un nombre de scénarios fixe, plus la dimension augmente, plus la distance moyenne entre les scénarios

Tableau 5.3 Distances de Wasserstein obtenues par la méthode des copules et la QVAC_1.

Instance			**Méthode**	
D	K	σ_{ji}	Copule	QVAC_1
2	50	0,3	0,375	0,293
2	50	0,8	0,393	0,230
2	250	0,3	0,179	0,135
2	250	0,8	0,183	0,109
10	50	0,3	7,176	5,199
10	50	0,8	7,147	3,072
10	250	0,3	5,870	4,528
10	250	0,8	5,953	2,684

et la médiane de la distribution diminue. En autres mots, les scénarios «se rapprochent» de la médiane. Par conséquent, les scénarios marginaux trouvés en 1^e étape correspondent aux valeurs optimales pour le cas unidimensionnel, mais sont trop éloignés de la médiane pour discrétiser adéquatement la distribution multinomiale.

En 2 dimensions, on aurait également pu utiliser l'heuristique de quadrillage, même s'il y a présence de corrélations entre les distributions marginales. Par contre, on peut constater les limites de cette méthode puisque la génération de seulement 2 centres en 10 dimensions induit un arbre de $2^{10} = 1024$ scénarios ; une quantité trop élevée pour la majorité des PS. Il faudrait donc s'en tenir à une unique réalisation pour certaines dimensions si l'on veut générer des arbres de taille plus modeste. Étant donné les résultats ci-dessus, on ne considérera pas la méthode des copules pour générer nos scénarios dans les tests à venir.

Analyse par composantes principales

Nous avons vu que l'heuristique de quadrillage peut entraîner des valeurs de $d_1(\mathbf{U}, \mathbf{V})$ assez près de celles obtenues par la QVAC lorsque la dimension et les corrélations sont assez faibles. Elle ne permet toutefois pas de considérer les corrélations et nécessite une très grande quantité de scénarios en dimension élevée. Nous avons alors analysé la méthode des copules qui permet de tenir compte des distributions marginales et de leurs corrélations séparément. On trouve cependant encore une fois que les résultats de cette procédure sont beaucoup moins bons que ceux de la QVAC. Nous tentons ici une dernière méthode appelée l'analyse par composantes principales (ACP), qui permet de réduire la dimensionnalité effective des données tout en préservant la majorité de la variance contenue dans le problème.

Présentons brièvement le concept de l'ACP. Soit $\mathbf{U} \in \mathbb{R}^D$ un vecteur aléatoire avec matrice des covariances $\Sigma_{\mathbf{U}}$ que l'on souhaite discrétiser [1]. Le principe de l'ACP est de chercher un premier vecteur normalisé $\boldsymbol{\alpha}_1$ tel que

$$\mathrm{Z}_1 = \boldsymbol{\alpha}_1 \cdot \mathbf{U} = \sum_{i=1}^{D} \alpha_{1i} \cdot \mathrm{U}_i \tag{5.13}$$

possède une variance maximale. Pour empêcher $\boldsymbol{\alpha}_1$ de prendre des valeurs arbitrairement grandes et ainsi obtenir une variance de Z_1 infinie, on ajoute la contrainte suivante :

$$\|\boldsymbol{\alpha}_1\|^2 = \boldsymbol{\alpha}_1 \cdot \boldsymbol{\alpha}_1 = 1 \ , \tag{5.14}$$

c'est-à-dire que $\boldsymbol{\alpha}_1$ doit être normalisé. On cherche ensuite une nouvelle variable $\mathrm{Z}_2 = \boldsymbol{\alpha}_2 \cdot \mathbf{U}$ de variance maximale linéairement indépendante de Z_1, où $\boldsymbol{\alpha}_2$ est normalisé. On procède ainsi de suite jusqu'à ce qu'on obtienne des variables Z_m, $m = 1, ..., M$ linéairement indépendantes et de variances maximales. La variable Z_m est nommée la m^{e} composante principale de $\mathbf{U}$ et on appellera $\boldsymbol{\alpha}_m$ le m^{e} vecteur des poids. On espère ainsi pouvoir représenter le problème par un vecteur aléatoire $\mathbf{Z} = (\mathrm{Z}_1,...,\mathrm{Z}_M)$ conservant une grande proportion de la variance avec $M \ll D$.

On pourrait démontrer que les vecteurs de poids $\boldsymbol{\alpha}_m$ correspondent en fait aux vecteurs propres de $\Sigma_{\mathbf{U}}$ dont les valeurs propres λ_m sont données par

$$\lambda_m = Var[\mathrm{Z}_m] \ . \tag{5.15}$$

Ainsi, $\boldsymbol{\alpha}_1$ correspond au vecteur propre possédant la plus grande valeur propre, notée λ_1. Sous notation matricielle, la relation entre $\mathbf{Z}$ et $\mathbf{U}$ est

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_D \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1D} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{D1} & \alpha_{D2} & \cdots & \alpha_{DD} \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_D \end{pmatrix} \ . \tag{5.16}$$

On trouve que les vecteurs de poids sont également linéairement indépendants. Ils sont donc orthonormés et on peut écrire $AA^T = 1$, où A est la matrice dont les lignes correspondent

1. Nous avons supposé dans ce chapitre que la matrice des covariances était connue, mais l'ACP est aussi possible à partir d'un ensemble de données duquel on estime les covariances.

aux vecteurs de poids. On a donc

$$\mathbf{Z} = A\mathbf{U} \tag{5.17}$$

$$\mathbf{U} = A^T\mathbf{Z} \ . \tag{5.18}$$

Il pourrait être démontré que la matrice des covariances peut être décomposée en contributions de chaque composante principale par la relation suivante :

$$\Sigma_{\mathbf{U}} = \sum_{m=1}^{M} \lambda_m \ \boldsymbol{\alpha}_m \boldsymbol{\alpha}_m^T. \tag{5.19}$$

En observant les éléments sur la diagonale, on remarque d'abord que

$$Var[\mathrm{U}_i] = \sum_{m=1}^{M} \lambda_m \ \alpha_{mi}^2 \ . \tag{5.20}$$

De plus, (5.19) implique que la contribution de chaque terme $\lambda_m \ \boldsymbol{\alpha}_m \boldsymbol{\alpha}_m^T$ à la matrice des covariances tend à décroître avec m, puisque les valeurs propres λ_m sont classées en ordre décroissant et les vecteurs $\boldsymbol{\alpha}_m$ sont normalisés. En autres mots, l'ACP préserve non seulement la majorité de la variance, mais également une bonne partie des covariances. On invite le lecteur intéressé à consulter Jolliffe (2002) pour plus de détails.

La génération de scénarios à l'aide de L'ACP s'effectue en suivant les étapes suivantes :

1. Trouver toutes les composantes principales Z_m, $m = 1, ..., D$ de $\mathbf{U}$ à partir de la matrice des covariances $\Sigma_{\mathbf{U}}$
2. Sélectionner un sous-ensemble de composantes principales $\{\mathrm{Z}_m : m = 1, ..., M\}$
3. Générer des scénarios $\mathbf{y}_j$, $j = 1, ..., K$ pour le vecteur aléatoire $\mathbf{Z} = (\mathrm{Z}_1,...,\mathrm{Z}_M)$
4. Transformer les scénarios $\mathbf{y}_j$ de $\mathbf{Z}$ en scénarios $\mathbf{v}_j = A^T\mathbf{y}_j$ de $\mathbf{U}$

Même si les composantes principales Z_m ne sont pas corrélées par définition, elles ne sont pas indépendantes. Il est donc préférable de générer les scénarios de $\mathbf{Z}$ en considérant les M dimensions à la fois. En fait, même lorsque les distributions marginales sont indépendantes, on a vu que la QVAC offrait de meilleurs résultats que l'heuristique de quadrillage. L'ACP nous fournit les axes selon lesquels la variance est la plus élevée. Puisque la variance mesure l'espérance du *carré* de la distance entre une variable aléatoire et sa moyenne, il est plus logique de générer les scénarios en utilisant la QVAC avec norme L_2 lorsque l'on utilise l'ACP. Les méthodes de génération de scénarios seront notées QVAC_2 lorsque l'analyse par

composantes principales n'est pas utilisée et ACP lorsqu'elle l'est.

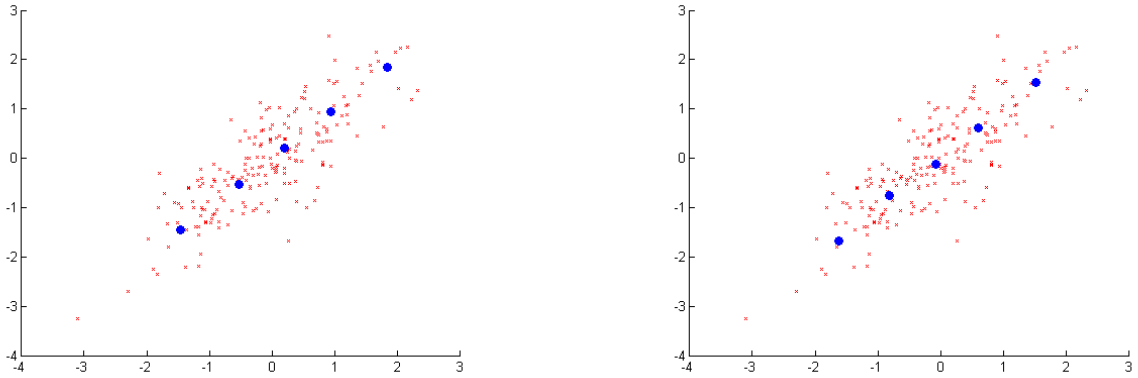

Figure 5.8 Scénarios obtenus avec la première composante principale de la matrice des covariances (figure de gauche). Scénarios obtenus par la QVAC (figure de droite).

La figure 5.8 représente 5 scénarios obtenus par la discrétisation d'un vecteur multinormale $\mathbf{U}$ avec moyennes et matrice des covariances

$$\boldsymbol{\mu} = (0,0) \qquad \text{et} \qquad \Sigma_{\mathbf{U}} = \begin{pmatrix} 1 & 0,8 \\ 0,8 & 1 \end{pmatrix} \tag{5.21}$$

en utilisant la première composante principale de $\mathbf{U}$ seulement. La variable Z_1 explique 90 % de la variance totale de $\mathbf{U}$. On observe que les scénarios sont alignés le long de la droite supportant le vecteur des poids $\boldsymbol{\alpha}_1 = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$. Des réalisations de $\mathbf{U}$ été ajoutées au graphique pour permettre de mieux évaluer la qualité des scénarios visuellement. La figure 5.8 représente les scénarios obtenus par la QVAC_2. On constate remarquablement que ceux-ci tendent également à s'aligner le long de la même droite. En fait, on obtient des distances de Wasserstein très similaires qui oscillent autour de 0,348 pour la QVAC_2 et lorsqu'on utilise l'ACP avec la 1^e composante principale.

Évidemment, l'ACP a produit un excellent résultat pour cet exemple puisque les distributions marginales de $\mathbf{U}$ étaient très corrélées. Notons toutefois qu'il est permis d'utiliser autant de composantes que l'on désire. Ainsi, si les distributions marginales de $\mathbf{U}$ avaient été peu corrélées, il aurait été préférable d'utiliser les 2 premières composantes principales du problème, ce qui revient à faire subir une rotation à l'espace échantillonnal. Du point de vue de la génération de scénarios avec la norme L_2, ces deux problèmes sont équivalents. Il s'agit ici d'une raison supplémentaire de ne pas utiliser la norme L_1 avec l'ACP, puisque la mini-

misation de la distance de Wasserstein n'est alors plus invariante par une rotation de l'espace.

Nous analysons maintenant la génération de scénarios à l'aide de l'ACP pour les cas qui nous intéressent, c'est-à-dire lorsque la dimension est élevée. Nous avons commencé par estimer une matrice des covariances en 25 dimensions à partir de données concernant la demande en employés pour un restaurateur. Ces données nous ont été transmises par l'entreprise, mais la signature d'un contrat de confidentialité nous empêche de les divulguer. La moyenne de la valeur absolue des corrélations estimées est

$$\frac{1}{D(D-1)} \sum_{i,j \ : \ i \neq j} \frac{|Cov[U_i, U_j]|}{(Var[U_i] \ Var[U_j])^{1/2}} = 0,346 \ . \tag{5.22}$$

On suppose ensuite que l'on cherche à discrétiser à l'aide de 50 scénarios une variable aléatoire multinomiale $\mathbf{U}$ dont le vecteur moyen est $\boldsymbol{\mu} = \mathbf{0}_{25} = (0, ..., 0)$ et possédant la matrice des covariances estimée. Une question fondamentale de l'ACP consiste à déterminer la quantité M de composantes principales que l'on devrait utiliser. Nous choisirons un nombre de composantes principales qui minimisent la distance de Wasserstein pour l'ACP. La courbe de la $d_1(\mathbf{U}, \mathbf{V})$ en fonction de M est représentée à la figure 5.9.

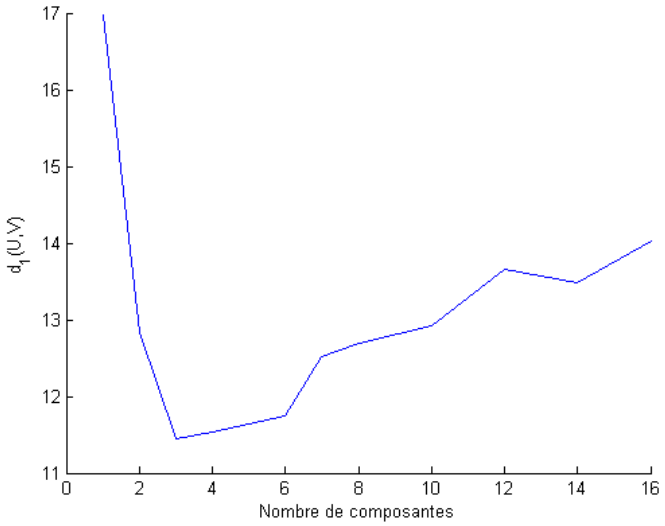

Figure 5.9 Distance de Wasserstein en fonction du nombre de composantes principales retenues

On constate que le minimum est atteint lorsque seulement 3 composantes principales sur une possibilité de 25 sont utilisées, représentant 71,07 % de la variance totale de $\mathbf{U}$. La proportion

de la variance considérée augmente avec M, mais le problème de minimisation de la distance de Wasserstein comporte alors plus de dimensions et devient plus difficile à résoudre. Une composante principale doit donc contribuer significativement à l'explication de la variance pour mériter d'être utilisée.

Le tableau 5.4 compare les distances de Wasserstein obtenues par la QVAC_2 avec celles résultant de l'ACP pour différentes instances. La matrice des covariances en 10 dimensions a été estimée à partir du même ensemble de données que pour $D = 25$. Le pourcentage de la variance expliquée $e_{\%}$ est affiché dans la 4^e colonne et le gain relatif (voir l'équation 5.6) de l'ACP par rapport à la QVAC_2 dans la dernière.

Tableau 5.4 Distances de Wasserstein obtenues avec l'analyse par composantes principales (ACP) et la QVAC_2.

Instance				**Méthode**		**Gain relatif**
D	M	K	$e_{\%}$ (%)	ACP	QVAC_2	$G_{\%}$ (%)
10	4	50	86,23	2,553	2,900	11,96
10	5	50	90,66	2,624	2,900	9,52
10	4	250	86,23	1,849	2,049	9,76
10	5	250	90,66	1,800	2,049	12,15
25	3	50	71,07	11,514	14,265	19,28
25	5	50	82,67	11,621	14,265	18,08
25	8	50	90,58	12,424	14,265	12,91
25	3	250	71,07	9,807	10,954	10,47
25	5	250	82,67	8,637	10,954	21,15
25	8	250	90,58	9,191	10,954	16,09

On remarque d'abord que la génération de scénarios par l'ACP engendre des valeurs de $d_1(\mathbf{U},\mathbf{V})$ considérablement inférieures à la QVAC_2 pour toutes les instances. Comme on pouvait s'y attendre, le gain relatif est plus important lorsque la dimension augmente. Notons également que le nombre de composantes principales optimal dépend de la dimension du problème [2] et de la quantité K de scénarios générés. Par exemple, les meilleurs résultats avec 50 scénarios en 10 et 25 dimensions sont obtenus lorsque $M = 3$ et $M = 4$ respectivement. Si on utilise plutôt 250 scénarios, il est optimal d'utiliser 5 composantes principales dans les deux cas. On remarque également que si on utilise de trop de composantes principales ($M = 8$ par exemple), alors on n'atteint pas d'aussi bons résultats. Avec le bon choix de M, le gain

2. En fait, la matrice des covariances ne serait plus la même pour 2 problèmes de dimensions différentes.

relatif est environ le même indépendamment du nombre de scénarios.

On conclut des résultats ci-dessus que parmi les trois méthodes de réduction de la dimension, seule l'analyse par composantes principales permet de réduire la distance de Wasserstein obtenue par la QVAC_2 en dimension élevée. En fait, la génération de scénarios multidimensionnels devrait toujours se faire à l'aide de l'ACP lorsqu'il y a présence de corrélations entre les distributions marginales. Les résultats ne peuvent être pires que ceux de la QVAC_2 puisqu'il est possible dans des cas extrêmes de se servir de toutes les composantes principales et ainsi obtenir une génération de scénarios équivalente à celle de la QVAC_2.

5.1.3 Problème analytique

La propriété la plus importante d'une méthode de génération de scénarios reste néanmoins de produire une solution aussi près que possible de la solution optimale, c'est-à-dire que $e(Z, \tilde{Z}) \approx 0$. La distance de Wasserstein permet d'évaluer une borne supérieure sur l'erreur commise par la discrétisation de $\mathbf{U}$, mais ne nous permet pas de la mesurer directement. Comme nous l'avons déjà mentionné, l'évaluation de $e(Z, \tilde{Z})$ n'est pas toujours possible puisqu'elle requiert la connaissance de la solution optimale réelle qui serait obtenue si l'on pouvait résoudre le PS à partir de la vraie distribution. On peut parfois obtenir une valeur approximative de l'erreur en résolvant le PS avec un arbre de scénarios de référence aussi gros que possible, mais cela n'est pas toujours faisable. Pour réellement tester les scénarios, nous avons besoin d'un problème d'optimisation stochastique que l'on peut résoudre analytiquement.

Cas unidimensionnel

Nous avons donc choisi d'évaluer les scénarios à l'aide du fameux problème du vendeur de journaux (*newsvendor problem*), qui s'énonce comme suit. Un vendeur achète et vend des journaux d'un fournisseur à des prix unitaires c et b respectivement ($c < b$). Les journaux invendus sont rachetés par le fournisseur à la fin de la journée au prix unitaire r ($< c$). On suppose que la demande des clients est incertaine et suit une distribution U. Le problème consiste à déterminer la quantité de journaux optimale à acheter en fonction de la demande des clients. On notera x, y et w respectivement les quantités de journaux achetés, vendus et rachetés par le fournisseur. Ce problème est typiquement utilisé pour déterminer la quantité optimale d'un produit périssable que l'on devrait détenir en stock. Le problème du vendeur de journaux est l'un des rares que l'on est en mesure de résoudre exactement. Il nous permettra donc de mesurer l'erreur d'approximation donnée par (3.8).

En suivant la notation de l'équation (3.5), le programme stochastique se formule donc comme suit :

$$Z(x) = \min_{x} \ \ cx + E_{\mathrm{U}}[Q(x, \mathrm{U})] \ , \tag{5.23}$$

où

$$\begin{aligned} Q(x, \mathrm{U}) = \min_{y,w} \quad & -by - rw \\ \text{s.c.} \quad & y \leq \mathrm{U} \\ & y + w \leq x \\ & y, \ w \geq 0 \end{aligned} \tag{5.24}$$

et $b, r, c \geq 0$. Il est bien connu (voir Birge et Louveaux (2011), par exemple) que les valeurs optimales du problème de 2^e étape sont

$$y^* = \min(x, \mathrm{U}) \tag{5.25}$$

$$w^* = \max(x - \mathrm{U}, 0) \tag{5.26}$$

et on trouve que la solution optimale est

$$x^* = \begin{cases} 0 & si \ \frac{b-c}{b-r} < F_{\mathrm{U}}(0) \\ F_{\mathrm{U}}^{-1}\left(\frac{b-c}{b-r}\right) & sinon \ , \end{cases} \tag{5.27}$$

où F_{U} est la fonction de répartition de U. On a donc

$$\begin{aligned} Z(x^*) &= cx^* + E_{\mathrm{U}}[-b \ \min\{x^*, \mathrm{U}\} - r \ \ \max\{x^* - \mathrm{U}, 0\}] \\ &= cx^* + \int_{-\infty}^{x^*} (-bu - r(x^* - u)) dF_{\mathrm{U}}(u) + \int_{x^*}^{\infty} -bx^* dF_U(u) \\ &= cx^* - rx^* F_{\mathrm{U}}(x^*) + (r - b) \int_{-\infty}^{x^*} u \ dF_{\mathrm{U}}(u) - bx^*(1 - F_{\mathrm{U}}(x^*)) \\ &= cx^* - rx^* F_{\mathrm{U}}(x^*) + (r - b) \left(x^* \ F_{\mathrm{U}}(x^*) - \int_{-\infty}^{x^*} F_{\mathrm{U}}(u) du \right) - bx^*(1 - F_{\mathrm{U}}(x^*)) \\ &= (c - b)x^* + (b - r) \int_{-\infty}^{x^*} F_{\mathrm{U}}(u) du \ , \end{aligned} \tag{5.28}$$

où l'on a intégré par parties à l'avant-dernière ligne.

On démontre maintenant que la fonction de coût est uniformément Lipschitz de constante b. Notons tout d'abord que

$$z(x, \mathrm{U}) = cx - b \min\{x, \mathrm{U}\} - r \; max\{x - \mathrm{U}, 0\} \; , \tag{5.29}$$

puisque

$$\begin{aligned} Z(x) &= \min_x \; cx + E_{\mathrm{U}}[-b \min\{x, \mathrm{U}\} - r \;\; \max\{x - \mathrm{U}, 0\}] \\ &= \min_x \; E_{\mathrm{U}}[\; cx - b \min\{x, \mathrm{U}\} - r \;\; \max\{x - \mathrm{U}, 0\}] \\ &= \min_x \; E_{\mathrm{U}}[z(x, u)] \; . \end{aligned}$$

On a donc

$$\begin{aligned} |z(x, u) - z(x, v)| = | \; & (cx - b \min\{x, u\} - r \;\; \max\{x - u, 0\}) \\ & - (cx - b \min\{x, v\} - r \;\; \max\{x - v, 0\}) \; | \\ = |b \; & (\min\{x, v\} - \min\{x, u\}) + r \; (\max\{x - v, 0\} - \max\{x - u, 0\})| \; . \end{aligned}$$

En supposant sans perte de généralité que $v > u$, on a

$$\begin{aligned} 0 \leq (\min\{x, v\} - \min\{x, u\}) \leq (v - u) \\ -(v - u) \leq (\max\{x - v, 0\} - \max\{x - u, 0\}) \leq 0 \end{aligned}$$

et puisque $b > r$,

$$\begin{aligned} |z(x, u) - z(x, v)| &\leq | \; b \; (v - u) + 0| \\ &= b \; |v - u| \; . \end{aligned}$$

La fonction de coût est donc uniformément Lipschitz de constante $\overline{L}_1 = b$. La relation 3.16 nous donne donc la borne supérieure sur l'erreur d'approximation pour ce PS :

$$e(Z, \tilde{Z}) \leq 2b \cdot d_1(\mathrm{U}, \mathrm{V}) \; . \tag{5.30}$$

Puisque la demande ne peut être négative, on supposera que celle-ci suit une loi log-normale

$$\mathrm{U} \sim \text{Log-}\mathcal{N}(\mu, \sigma^2) \; . \tag{5.31}$$

C'est-à-dire que $\log(\mathrm{U})$ suit une loi normale de moyenne μ et variance σ^2. On a choisi $\mu =$

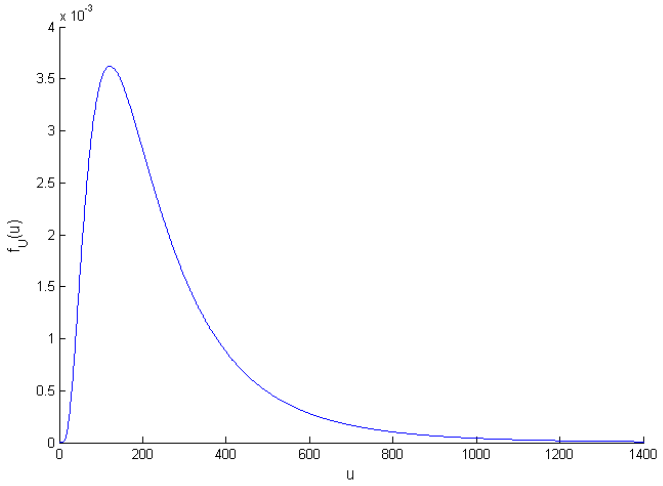

Figure 5.10 Fonction de densité d'une variable Log-$\mathcal{N}(\log(200), 1/2)$.

$\log(200)$ et $\sigma^2 = 1/2$. La fonction de densité est représentée sur la figure 5.10. On analysera le problème du vendeur de journaux avec $b = 5$, $c = 2$ et $r = 1$. Avec ces valeurs des paramètres, on calcule la solution et le coût optimal pour $\mathrm{U} \sim \text{Log-}\mathcal{N}(\log(200), 1/2)$:

$$\begin{aligned} x^* &= F_{\mathrm{U}}^{-1}\left(\frac{b-c}{b-r}\right) \\ &= F_{\mathrm{U}}^{-1}(3/4) \approx 322,22 \end{aligned}$$

et

$$\begin{aligned} Z(x^*) &= -3\ x^* + 4 \int_0^{x^*} F_{\mathrm{U}}(u)du \\ &= -3 \times 322,22 + 4 \int_0^{322,22} \left(\frac{1}{2} + \frac{1}{2}\ erf\left(\frac{\log(u) - \log(200)}{\sqrt{2 \cdot 1/2}}\right)\right) du \\ &= -966,66 + 4 \times 116,60 \\ &= -500,25\ , \end{aligned} \tag{5.32}$$

où erf est la fonction d'erreur.

La figure 5.11 représente la discrétisation de U pour 50 scénarios par l'algorithme de Lloyd_1. La représentation des événements extrémaux est souvent une qualité recherchée des arbres de scénarios. On remarque effectivement que la discrétisation de U par la méthode de Lloyd_1 représentée sur la figure 5.11 couvre assez bien les réalisations extrêmes en incluant 5

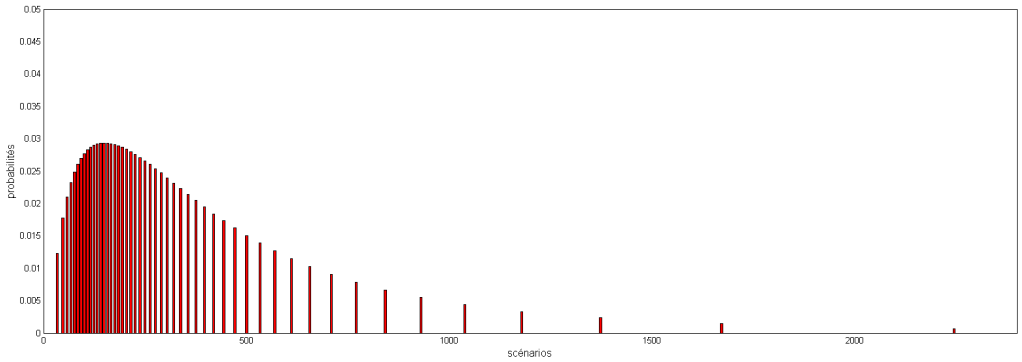

Figure 5.11 Représentation de 50 scénarios obtenus par l'algorithme de Lloyd$_1$ discrétisant une variable Log-$\mathcal{N}(\log(200), 1/2)$

scénarios dont la valeur est supérieure à 1000. Pour cette quantification, on trouve $d_1(\mathrm{U}, \mathrm{V}) = 4,711$. L'inégalité 5.30 nous indique donc que

$$err(Z, \tilde{Z}) \leq 2 \times 5 \times 4,711 = 47,11 \ . \tag{5.33}$$

Le coût de la solution obtenue par les scénarios sera donc inférieur ou égal à 47,11 unités de plus que la solution optimale réelle, i.e. $Z(\tilde{x}^*) \leq -500,25 + 47,11 = -453,14$. La solution optimale mène donc à un profit moyen de $500,25$ unités, tandis la solution approximative obtenue par la discrétisation donnera un profit moyen d'au moins 453,14.

Une fois les scénarios générés, la discrétisation du PS pour le problème du vendeur de journaux s'écrit comme suit :

$$Z(x) = \min_x \ \ cx + \sum_{j=1}^{K} p_j \ [Q(x, v_j)] \ , \tag{5.34}$$

où $Q(x, v_j)$ est définie par (5.24). On trouve en fait que la quantification représentée à la figure 5.11 mène à la solution $\tilde{x}^* = 317,301$, ce qui donne $Z(\tilde{x}^*) = -500,178$. L'erreur obtenue est donc égale à 0,072 et est nettement inférieure à la borne supérieure calculée.

Si l'on utilise plutôt l'algorithme de Lloyd$_2$ (avec la norme L_2), on obtient les scénarios représentés sur la figure 5.12. Nous avons choisi la même échelle graphique pour les figures 5.11 et 5.12 de manière à pouvoir comparer les discrétisations obtenues. Il y a toutefois 4 scénarios obtenus par l'algorithme de Lloyd$_2$ situés approximativement aux points 2629, 3136, 3950 et

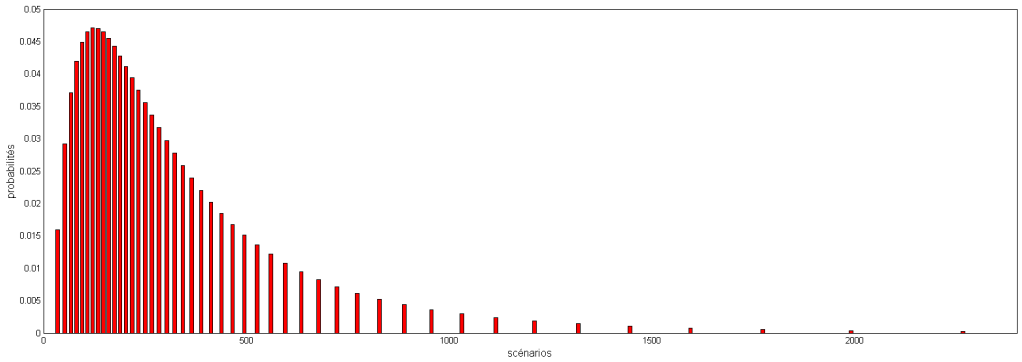

Figure 5.12 Représentation de 50 scénarios obtenus par l'algorithme de Lloyd$_2$ discrétisant une variable Log-$\mathcal{N}(\log(200), 1/2)$

5550 qui n'ont pu être représentés graphiquement en conséquence. Incluant ces 4 scénarios, on obtient un total de 5 valeurs supérieures à 2000, qui possèdent tous évidemment une très faible probabilité de réalisation. On constate que les événements extrêmes sont encore mieux représentés en utilisant la norme L_2, mais réserve moins de scénarios pour les événements plus probables. Cette dispersion des scénarios résulte en de plus grandes probabilités de réalisation pour ceux situés près des modes. On en déduit que le choix de la norme dépend du problème et de la représentation des événements extrémaux désirée.

Les valeurs des solutions obtenues par l'algorithme de Lloyd (en norme L_1 et L_2) sur le problème du vendeur de journaux sont présentées dans le tableau 5.5. Elles sont comparées à celles de l'échantillonnage pur et de l'optimisation déterministe où l'unique scénario correspond au vecteur des moyennes marginales de $\mathbf{U}$, c'est-à-dire avec un seul scénario $\mathbf{v} = (E[\mathrm{U}_1],\ E[\mathrm{U}_2], ..., E[\mathrm{U}_D])$ de probabilité 1. Un exemple de discrétisation obtenu par échantillonnage pur est représenté à la figure 5.13. On constate que les scénarios sont placés de manière beaucoup plus désorganisée que pour les algorithmes de Lloyd.

Avant d'analyser le tableau 5.5, il faut savoir que la solution obtenue est très variable pour une même méthode d'une implémentation à l'autre. Ceci est particulièrement vrai pour l'échantillonnage pur. Le tableau 5.5 affiche donc la valeur moyenne de $Z(\tilde{x}^*)$ sur 10 implémentations de l'échantillonnage pur. Notons également que les valeurs des solutions auraient été différentes pour un autre choix des paramètres b, c et r ou pour un PS carrément différent. Les résultats présentés nous permettront tout de même de tirer certaines conclusions intéressantes par rapport aux méthodes de génération de scénarios.

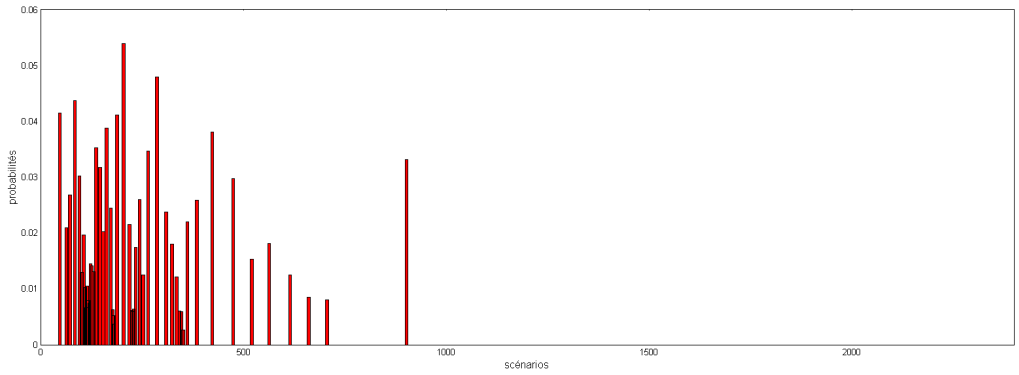

Figure 5.13 Représentation de 50 scénarios obtenus par échantillonnage pur discrétisant une variable Log-$\mathcal{N}(\log(200), 1/2)$.

Tableau 5.5 Solutions des algorithmes de Lloyd_1 et Lloyd_2, de l'échantillonnage pur (EP) et de l'optimisation déterministe (OD)

		$-Z(\tilde{x}^*)$				$-Z(x^*)$
D	K	Lloyd_1	Lloyd_2	EP	OD	sol. exacte
1	5	499,00	494,50	489,93		
1	50	500,18	500,18	499,66	486,57	500,25
1	100	500,21	500,22	500,16		

On remarque d'abord que les algorithmes de Lloyd_1 et Lloyd_2 induisent des gains considérablement supérieurs à ceux de l'optimisation déterministe et l'échantillonnage pur. Pour $K = 5$, la solution fournie par l'algorithme de Lloyd est meilleure avec la norme L_1 qu'avec L_2, mais la différence devient insignifiante à partir de $K = 50$. Avec 100 scénarios, on obtient par les méthodes de Lloyd des valeurs des solutions supérieures à 99,99 % le gain optimal. L'échantillonnage pur offre également d'excellents résultats avec 100 scénarios et conduit à des valeurs de $Z(\tilde{x}^*)$ très près de l'optimalité. Comme on pouvait s'y attendre, on constate que l'optimisation stochastique est plus profitable par rapport à l'optimisation déterministe pour le problème du vendeur de journaux lorsqu'une grande quantité de scénarios sont utilisés. D'autre part, les gains obtenus par les méthodes de Lloyd relativement à l'échantillonnage pur sont plus significatifs lorsque peu de scénarios sont utilisés.

Cas multidimensionnel

On considère maintenant une version multidimensionnelle du problème du vendeur de journaux qui servira à tester les algorithmes de QVAC lorsque la dimension est supérieure à 1. On suppose que notre vendeur souhaite acheter et vendre D biens dont la distribution de la demande est U_i, $i = 1, ..., D$ à des prix c_i et b_i respectivement. Le prix de rachat est r_i et les variables de décision analogues au problème ci-dessus seront notées x_i, y_i et w_i. Le PS s'énonce donc comme suit :

$$Z(x) = \min_x \ \sum_{i=1}^{D} c_i x_i + E_{\mathrm{U}_i}[Q(x_i, \mathrm{U}_i)] \ , \tag{5.35}$$

où $Q(x, v_j)$ est définie par (5.24). Le PS ci-dessus peut être divisé en D sous-problèmes indépendants qui pourraient être résolus séparément. Autrement dit, il est donc équivalent à D programmes stochastiques unidimensionnels. On n'aurait alors qu'à générer les scénarios en discrétisant chaque distribution U_i individuellement par l'algorithme de Lloyd de manière à obtenir des solutions environ D fois supérieures à celles trouvées en une dimension. Cependant, le but de l'exercice ici est de tester les méthodes de génération de scénarios pour le cas multidimensionnel. Nous procéderons donc à la quantification du vecteur $\mathbf{U}$ en entier.

On considérera pour commencer que les variables aléatoires sont indépendantes et suivent une loi $\mathrm{U}_i \sim \text{Log-}\mathcal{N}(\log 200, 1/2)$ avec $b_i = 5$, $c_i = 2$ et $r_i = 1 \ \forall \ i$. Notons que si les valeurs de ces paramètres avaient été différentes pour certaines dimensions, alors les distributions marginales n'auraient pas toutes eu la même influence sur le PS et il aurait été préférable d'appliquer la transformation Γ introduite à la proposition 3.4 *a priori*. Comme nous l'avons vu précédemment, la génération de scénarios n'est pas triviale même lorsque les variables ne sont pas corrélées. Il est important de considérer cette situation, puisqu'il n'est certainement pas avantageux d'utiliser l'ACP lorsque les variables sont indépendantes. On se servira donc directement des algorithmes QVAC pour générer les scénarios.

La variance des distributions marginales U est donnée par

$$\begin{aligned} Var[\mathrm{U}] &= (e^{\sigma^2} - 1)\ e^{2\mu+\sigma^2} \\ &= (e^{1/2} - 1)\ e^{2\ (200)+1/2} \\ &\approx 42\ 782\ . \end{aligned}$$

On utilisera donc la QVAC_1 avec

$$\gamma_n = 0.8 \times \sqrt{42\ 782} \times \frac{K}{K+n}\ . \tag{5.36}$$

Avec la méthode de QVAC_1 tel que présenté au chapitre précédent, si la valeur d'un centre v_{ji}^n se situe près de 0 à une itération n donnée et qu'on tire un échantillon inférieur à cette valeur, il se pourrait alors qu'elle «dépasse» l'axe des ordonnées et prenne une valeur inférieure à 0. Puisqu'une loi log-normale ne peut prendre de valeurs négatives, on doit faire un ajustement particulier à la QVAC_1 pour empêcher les scénarios inférieurs à 0. Nous avons donc choisi d'attribuer une nouvelle valeur échantillonnée à partir de U à tout centre qui aurait été déplacé vers une valeur négative à l'itération n. Ce centre sera ensuite itéré comme tous les autres par la procédure normale de l'algorithme de QVAC_1. Un ajustement de la sorte est effectivement nécessaire pour toutes les distributions dont le domaine réalisable est borné inférieurement et/ou supérieurement, telles que la loi uniforme. Notons que les solutions présentées dans le tableau 5.2 ont été obtenues sans appliquer cet ajustement et pourraient donc être améliorées à l'aide de la procédure proposée ici. Un avantage de la QVAC_2 est que le paramètre de saut dépend de la distance entre v_{ji} et un échantillon de U_i. Il n'y a donc aucun risque d'obtenir des scénarios négatifs et l'algorithme ne requiert aucun ajustement particulier.

Nous avons remarqué durant les tests que les résultats obtenus en appliquant la QVAC 2 fois de suite étaient meilleurs que si on ne l'exécutait qu'une fois. Autrement dit, la première implémentation sert d'initialisation des centres pour la deuxième. Il n'est cependant pas nécessaire d'exécuter l'algorithme plus de 2 fois. Elle peut donc être utilisée avec un nombre d'échantillons limité puisqu'on ne cherche pas directement la solution optimale, mais seulement une initialisation raisonnable. Nous avons initialisé les centres avec 5×10^5 échantillons à partir desquels l'algorithme était implémenté une 2^e fois avec 10^7 tirages. Les valeurs de $Z(\tilde{x}^*)$ obtenues à partir des méthodes de QVAC, de l'échantillonnage pur et de l'optimisation déterministe se retrouvent dans le tableau 5.6. La norme L_2 a été utilisée pour calculer les probabilités de l'échantillonnage pur dont les résultats correspondent encore à une moyenne sur 10 implémentations.

Pour $D = 2$, les méthodes de QVAC mènent à de meilleurs résultats que l'échantillonnage pur et l'optimisation déterministe. On note cependant que la QVAC_1 semble donner des meilleurs résultats que la QVAC_2. Ceci ne doit probablement pas être considéré comme une généralité, mais plutôt comme une spécificité de notre problème du vendeur de journaux.

Tableau 5.6 Solutions de la QVAC_1, QVAC_2, de l'échantillonnage pur (EP) et de l'optimisation déterministe (OD)

		$-Z(\tilde{x}^*)$				$-Z(x^*)$
D	K	QVAC_1	QVAC_2	EP	OD	sol. exacte
2	5	998,45	989,70	975,79		
2	50	1000,17	999,47	996,25	963,14	1000,50
2	100	1000,24	1000,11	997,78		
10	50	4579,43	4911,49	4876,55		
10	100	4637,94	4949,04	4903,59	4865,69	5002,50
10	500	4929,51	4975,47	4952,79		

En 10 dimensions, on remarque que la QVAC_1 génère des résultats désastreux. L'explication de ce phénomène est en lien avec la tendance des scénarios générés avec la norme L_1 (resp. L_2) de se rapprocher de la médiane (resp. moyenne) lorsque la dimension augmente. Nous avions également mentionné cette caractéristique lors de l'analyse de la méthode des copules. Ainsi, les scénarios obtenus par la discrétisation de $\mathbf{U}$, dont les distributions marginales suivent une loi Log-$\mathcal{N}(\log(200), 1/2)$, avec la QVAC_1 se trouveront près du vecteur médian $\boldsymbol{\nu}^{1/2} = (200, ..., 200)^T$ tandis que ceux de la QVAC_2 se concentreront autour du vecteur moyen $\boldsymbol{\mu} \approx (257, ..., 257)^T$. Étant donné que la solution optimale du problème du vendeur de journaux avec notre choix des paramètres est $\mathbf{x}^* = F_{\text{U}}^{-1}(3/4) \approx (322, ..., 322)$, il n'est pas surprenant que les gains obtenus avec la norme L_2 soient supérieurs. À moins de considérer jusqu'à 500 scénarios, la QVAC_1 demeure encore pire que l'optimisation déterministe, où seul le vecteur moyen est considéré. La QVAC_2 est l'algorithme le plus efficace pour ce problème en 10 dimensions. Les coûts obtenus par cette méthode avec 50 et 100 scénarios sont similaires à ceux de l'échantillonnage pur avec 100 et 500 scénarios respectivement. Néanmoins, les coûts de la QVAC_2 et de l'échantillonnage pur ne sont pas si loin de celles de l'optimisation déterministe lorsque $K = 50$. En bref, il est clair qu'un grand nombre de scénarios doit être considéré pour que l'optimisation stochastique soit profitable en dimension élevée. Notons cependant que même si 500 scénarios peuvent paraître comme beaucoup, cette quantité représente environ la moitié de ceux qui seraient obtenus par la combinaison de 2 scénarios marginaux pour chacune des 10 dimensions ($2^{10} = 1024$).

On considère maintenant le problème du vendeur de journaux multidimensionnel (5.35) où les distributions marginales U_i suivent une loi log-normale de moyenne $\mu_i = \log 200$ et sont corrélées. On évaluera les méthodes de génération de scénarios sur un vecteur aléatoire $\mathbf{U}$ à 10 dimensions. À la sous-section sur l'ACP, on avait obtenu une matrice des covariances à

partir d'un ensemble de données confidentielles sur la demande en employés dans le domaine de la restauration. On s'en servira à nouveau pour calculer les covariances de $\mathrm{U}_i = e^{\mathrm{X}_i}$, où $\mathbf{X} = (\mathrm{X}_1, ..., \mathrm{X}_D)$ suit une loi multinormale avec la matrice des covariances trouvée précédemment. La moyenne de la valeur absolue des corrélations utilisées pour les variables U_i est 0,329. On suppose encore que $b_i = 5$, $c_i = 2$ et $r_i = 1 \; \forall \; i \in \{1, ..., D\}$.

Notons que le programme stochastique (5.35) ne dépend pas des corrélations entre les distributions marginales. Il peut donc être divisé en D sous-problèmes et il suffirait de discrétiser les distributions marginales individuellement. La qualité des solutions pour ce problème reflète donc la capacité des méthodes de génération de scénarios à bien représenter les distributions marginales, mais ne dépend pas de la représentation des corrélations. Il serait possible de créer des PS stochastiques dont les solutions dépendent des corrélations, mais ceux-ci ne peuvent généralement pas être résolus analytiquement comme on l'a fait pour le problème du vendeur de journaux. Pour remédier à cet inconvénient, on pourrait utiliser une très grande quantité de scénarios et ainsi trouver une solution assez près du minimum global, à partir de laquelle on compare les autres solutions obtenues. Or, ces scénarios doivent être générés par une méthode différente de celles qui sont testées pour éviter de biaiser les résultats (Kaut et Wallace (2003)). Supposons par exemple que la QVAC_2 est utilisée avec un grand nombre de scénarios pour obtenir une approximation de la solution optimale. Il y a alors plus de chances que les solutions obtenues par la QVAC_2 s'apparentent à l'estimation de la solution optimale que pour toute autre méthode testée. Par conséquent, il serait préférable d'approximer la solution optimale à partir de l'échantillonnage pur pour être en mesure de tester les méthodes de QVAC et d'ACP équitablement. Cependant, même avec 10 000 scénarios, l'EP mène à une erreur d'approximation $e(Z, \tilde{Z}) = 8,01$ pour le programme stochastique (5.35) lorsque les distributions marginales sont indépendantes, ce qui est beaucoup trop loin de la solution optimale. Il est donc fort probable qu'une solution obtenue par échantillonnage pur avec une grande quantité de scénarios ne puisse servir d'échelon pour tester nos méthodes avec un PS quelconque. Par conséquent, le programme stochastique (5.35) sera considéré pour évaluer la génération de scénarios sur des variables aléatoires dépendantes, même s'il ne permet pas d'évaluer la capacité des méthodes à bien estimer les corrélations.

Lorsque les distributions marginales sont indépendantes, nous avons vu qu'en prenant suffisamment de scénarios, les méthodes de QVAC conduisaient à des gains considérablement supérieurs à ceux de l'échantillonnage pur et de l'optimisation déterministe. On peut se convaincre qu'on en arriverait aux mêmes conclusions lorsque les variables sont corrélées. Nous avons donc seulement comparé les solutions de la QVAC_2 et de l'ACP pour le problème

décrit ci-dessus. Pour choisir le nombre de composantes principales M, nous avons calculé la distance de Wasserstein en fonction de M pour chaque instance puis avons choisi celle qui minimisait $d_1(\mathbf{U}, \mathbf{V})$, comme il a été fait à la figure 5.9. Avec 50 et 100 scénarios, on trouve que la distance de Wasserstein minimale est atteinte avec 6 composantes principales, qui expliquent 92,3 % de la variance. Lorsque $K = 500$, on trouve cependant que l'on devrait utiliser toutes les 10 composantes principales, ce qui revient à discrétiser $\mathbf{U}$ directement à partir de la QVAC_2. Nous avons donc également utilisé l'ACP avec 6 composantes principales lorsque 500 scénarios sont générés afin de comparer les deux méthodes. Les résultats se retrouvent dans le tableau 5.7.

Tableau 5.7 Solutions de la QVAC_2 et de l'ACP avec 6 composantes principales sur le problème du vendeur de journaux.

		$-Z(\tilde{x}^*)$		$-Z(x^*)$
D	K	QVAC_2	ACP	sol.exacte
10	50	4867,04	4873,02	
10	100	4908,37	4908,88	4950,45
10	500	4941,01	4940,28	

Avec 50 et 100 scénarios, on constate que l'ACP permet d'obtenir des gains modestes par rapport à la QVAC_2. Lorsque $K = 500$, on avait déterminé en se basant sur les valeurs des distances de Wasserstein obtenues qu'il serait préférable d'utiliser 10 composantes principales plutôt que 6. On observe donc de meilleures solutions de la QVAC_2 pour ce cas, en accord avec nos prédictions. Ces résultats nous permettent de conclure deux choses. Premièrement, on constate que l'ACP permet souvent d'améliorer les solutions de la QVAC lorsqu'il y a présence de corrélations entre les distributions marginales, particulièrement lorsque peu de scénarios sont utilisés. Deuxièmement, il y a effectivement un lien entre la distance de Wasserstein et les solutions obtenues par l'optimisation stochastique. Bien qu'une valeur de $d_1(\mathbf{U}, \mathbf{V})$ inférieure ne garantisse pas une meilleure solution, il semble clairement y avoir une tendance des scénarios minimisant la distance de Wasserstein à produire des solutions de meilleure qualité. Il est donc pertinent de choisir le nombre de composantes principales en fonction de la capacité de l'ACP à minimiser $d_1(\mathbf{U}, \mathbf{V})$.

5.2 Distribution inconnue

On considère maintenant les problèmes pour lesquels la distribution du vecteur aléatoire est inconnue. C'est-à-dire que l'on suppose posséder un ensemble de données historiques à

partir desquelles il est difficile d'estimer une loi probabiliste, soit à cause d'une quantité de données insuffisantes ou parce que la distribution possède beaucoup trop de paramètres. Le but de cette section est de comparer les algorithmes d'échange des centres et de Lloyd (discret), qui permettent de générer les scénarios directement à partir des données.

On commence par considérer le cas unidimensionnel. Nous analyserons les algorithmes de partitionnement de données à partir d'échantillons provenant de la distribution suivante :

$$\mathrm{U} \sim \mathbb{I}_1 \, \mathcal{N}(-1,1) + \mathbb{I}_2 \, \mathcal{G}(5,1) + \mathbb{I}_3 \, \left(5 + 2 \cdot \mathcal{B}(5,2)\right) \ , \tag{5.37}$$

où

- $\mathcal{N}(-1,1)$ est une loi normale avec $\mu = -1$ et $\sigma^2 = 1$
- $\mathcal{G}(5,1)$ est une loi gamma avec paramètres de forme $k = 5$ et d'échelle $\theta = 1$
- $\mathcal{B}(5,2)$ est une loi beta avec paramètres de forme $\alpha = 5$ et $\beta = 2$
- $\mathcal{P}$ est une loi de probabilité discrète qui prend les valeurs q =1, 2 ou 3 avec probabilités respectives $p_1 = 0,3,\ p_2 = 0,5$ et $p_3 = 0,2$.
-

$$\mathbb{I}_q = \begin{cases} 1 & si\ \mathcal{P} = q \\ 0 & sinon\ . \end{cases} \tag{5.38}$$

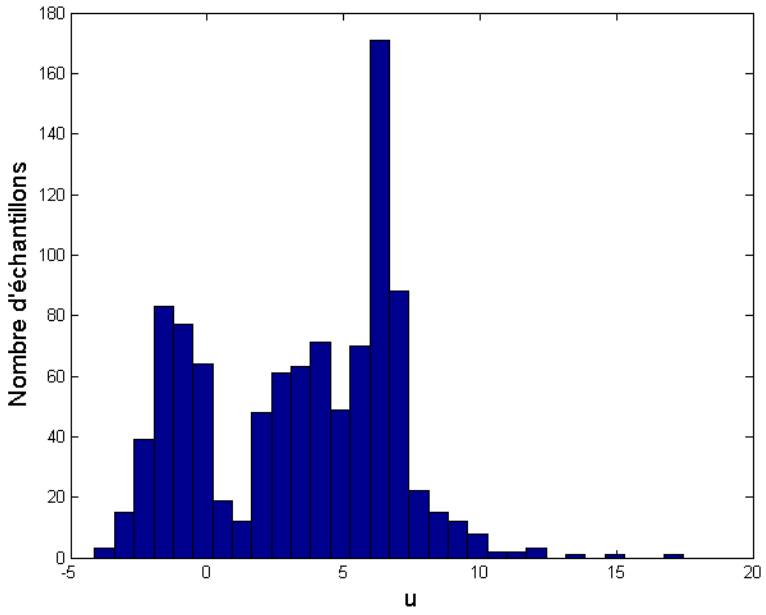

Figure 5.14 Histogramme de 1000 échantillons de U répartis dans 30 classes.

Nous avons choisi cette distribution pour son côté hétéroclite, ce qui la rend très difficile à estimer à l'aide d'un ensemble de données typiques. Un exemple d'histogramme avec 30 classes obtenu à partir de 1000 échantillons de U est représenté sur la figure 5.14. Notons toutefois que même s'il est pratiquement impossible d'obtenir la bonne forme de la distribution, il n'est pas clair que les scénarios devraient être générés directement à partir des données pour autant. Il semblerait également plausible que l'histogramme représenté à la figure 5.14 provienne d'une fonction de densité formée par la combinaison de trois gausiennes. Il est donc probable que l'on puisse trouver une régression acceptable de la distribution à l'aide des fonctions à base radiale, par exemple. Même si elle n'est pas exacte, la distribution ainsi estimée mènerait peut-être à des scénarios similaires à ceux obtenus à partir de la vraie loi de probabilité. Ce questionnement est en lien avec l'élaboration d'un critère permettant de déterminer s'il est préférable d'estimer la distribution ou non, que nous laissons ouvert à de futures recherches. La distribution énoncée ci-dessus nous permettra néanmoins de comparer les solutions obtenues à l'aide des algorithmes de Lloyd et d'échange des centres.

Dans le cas multidimensionnel, on supposera que les distributions marginales suivent toutes la même loi que U et sont non corrélées. Notons que si les variables étaient corrélées, il serait encore plus difficile de définir la distribution multinomiale puisqu'il faudrait estimer les corrélations en plus des paramètres des distributions marginales. Notre méthode sera testée sur des ensembles de 1000 échantillons de U.

Les résultats des algorithmes d'échange des centres et de Lloyd_1 (discret) sont présentés dans le tableau 5.8. On constate que le premier offre de meilleurs résultats lorsque la dimension est faible tandis que le second conduit à de meilleures solutions en dimension élevée. Il faut se rappeler que l'algorithme d'échange des centres traite le problème des k-médoïdes, où les centres sont contraints à être placés sur des points de données. Pour un problème à D dimensions, le choix d'un seul centre correspond à sélectionner D scénarios marginaux à la fois. Par conséquent, plus la dimension est élevée, moins bonnes sont les chances de trouver des centres parmi l'ensemble de données dont la valeur de chacune des dimensions discrétise adéquatement les distributions marginales. Peu importe la dimension, l'avantage d'un algorithme par rapport à l'autre se fait surtout remarquer lorsque le nombre de scénarios devient grand. On note que le temps de convergence de l'algorithme d'échange des centres est beaucoup plus long ($O(N^2D^2)$) que celui de Lloyd_1, mais reste tout de même raisonnable.

Nous avons mentionné au chapitre précédent que l'algorithme de Lloyd_1 générait une suite de solutions non croissantes et qu'il était donc possible de l'exécuter après l'algorithme

Tableau 5.8 Distance de Wasserstein et temps de convergence des algorithmes d'échange des centres (EC) et de Lloyd$_1$ (discret).

		EC		Lloyd$_1$	
D	K	$d_1(\mathbf{U},\mathbf{V})$	temps (s)	$d_1(\mathbf{U},\mathbf{V})$	temps (s)
1	5	0,5409	1,80	0,5653	0,005
1	50	0,0549	77,83	0,0853	0,011
1	100	0,0241	234,37	0,0437	0,018
2	5	2,255	1,50	2,321	0,009
2	50	0,682	80,3	0,712	0,022
2	100	0,446	210,90	0,459	0,026
10	5	24,735	1,83	23,991	0,058
10	50	17,051	65,82	15,712	0,071
10	100	14,767	235,40	12,999	0,132

d'échange des centres pour peaufiner les résultats. Nous avons cependant trouvé que le gain découlant de cette procédure est très faible et elle n'a donc pas été envisagée.

La figure 5.15 représente les 50 scénarios obtenus par l'algorithme d'échange des centres à partir des 1000 échantillons de U utilisés pour tester les méthodes de partitionnement de données en 2 dimensions. Il est aussi intéressant d'observer la capacité de cet algorithme à estimer les premiers moments marginaux de la distribution, considérant qu'il s'agit de l'unique objectif de certaines méthodes de génération de scénarios. Les espérances et variances marginales des 1000 données et 50 scénarios représentés sur la figure 5.15 sont affichées dans le tableau 5.9. On remarque que les moments des scénarios marginaux V_i sont assez près de ceux de l'ensemble de données U_i. L'algorithme d'échange des centres génère donc des arbres de scénarios qui respectent très bien les premiers moments de la distribution.

Tableau 5.9 Espérance et variances marginales des 1000 échantillons (U_i) et des 50 scénarios (V_i).

	U_1	V_1	U_2	V_2
E	3,56	3,53	3,41	3,36
Var	11,58	11,32	11,85	11,69

Nous avons à nouveau testé les scénarios générés sur le problème du vendeur de journaux, avec les mêmes paramètres que précédemment. La distribution considérée dans cette section peut prendre des valeurs négatives. Physiquement, ces réalisations ne sont pas réalistes puisqu'elles représentent une demande négative, mais le problème reste tout de même valide du

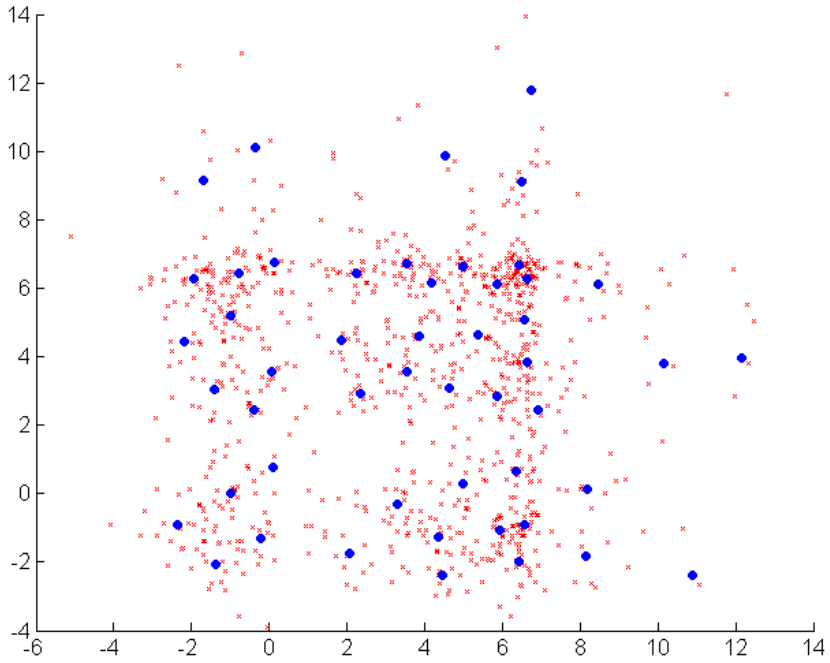

Figure 5.15 Représentation de 50 scénarios (points bleus) obtenus par l'algorithme d'échange des centres représentant 1000 données (croix rouges).

point de vue mathématique. On suppose que l'ensemble de 1000 données englobe la totalité des événements réalisables. Les coûts $Z(x^*)$ et $\tilde{Z}(\tilde{x}^*)$ correspondent donc respectivement aux programmes stochastiques résolus à partir de 1000 et K scénarios. Les résultats sont présentés dans le tableau 5.10. Les solutions exactes ont été obtenues en optimisant le problème du vendeur de journaux avec 1000 scénarios de probabilités égales, correspondant aux échantillons utilisés pour le partitionnement de données.

Sauf pour l'instance avec 5 scénarios en une dimension, on constate que les algorithmes de partitionnement de données engendrent des solutions plus près du minimum global que l'optimisation déterministe, où seul le vecteur de la moyenne des distributions marginales est considéré. Il ne semble pas y avoir une méthode nettement plus avantageuse entre l'algorithme d'échange des centres et celui de Lloyd$_1$. Sur certaines instances, la méthode de Lloyd$_1$ donne de meilleurs résultats que l'échange des centres et vice versa. Dans tous les cas, les résultats sont excellents lorsque 50 scénarios ou plus sont considérés. Nos méthodes de génération de scénarios nous permettent donc d'obtenir des solutions très près de l'optimalité avec seulement 50 scénarios représentant 1000 données. En comparant avec le tableau 5.8, on

Tableau 5.10 Coûts des solutions des algorithmes d'échange des centres (EC), de Lloyd_1 et de l'optimisation déterministe pour le problème du vendeur de journaux avec un ensemble de 1000 données.

		$-Z(\tilde{x}^*)$			$-Z(x^*)$
D	K	EC	Lloyd_1	OD	sol. exacte
1	5	4,0901	5,2590	5,2356	7,3159
1	50	7,2516	6,6881		
1	100	7,2541	7,3099		
2	5	13,9803	13,9911	10,0742	14,0122
2	50	14,0085	14,0054		
2	100	14,1030	14,0109		
10	5	61,1572	61,8168	44,9931	64,9726
10	50	64,8789	64,7160		
10	100	64,9232	64,8457		

remarque que même si la distance de Wasserstein obtenue par l'algorithme d'échange des centres était inférieure à celle de la méthode de Lloyd_1 en une dimension, nous avons obtenu de meilleurs résultats avec cette dernière lorsque $K = 5$ et $K = 100$. On en déduit que la minimisation de la distance de Wasserstein ne conduit pas nécessairement à une erreur d'approximation $e(Z, \tilde{Z})$ minimale. Étant donné la rapidité et la simplicité de l'algorithme de Lloyd, il est sans doute plus pratique de privilégier cette méthode par rapport à l'échange des centres dans la plupart des cas. Rappelons cependant que pour la discrétisation d'un vecteur aléatoire discret, il peut être préférable de choisir l'algorithme d'échange des centres pour que les scénarios obtenus correspondent à des points de données.

5.3 Application

Nous considérons maintenant le problème de confection d'horaires d'employés traité par Pacqueau (2011). L'objectif du programme stochastique est de minimiser les coûts liés au salaire des travailleurs d'une entreprise où le nombre d'employés requis à chaque instant de la journée est incertain. Celle-ci est divisée en 96 périodes de 15 minutes ; le vecteur aléatoire correspond donc à la demande en employés quotidienne et comporte 96 dimensions. Les différents ensembles contenus dans le problème sont présentés dans le tableau 5.11.

L'entreprise étudiée possède des employés à temps plein et à temps partiel, qui se distinguent par leur salaire et la longueur des quarts de travail. La fabrication des horaires de travail repose sur l'assignation des quarts aux employés à temps plein, à temps partiel, des pauses et des heures supplémentaires. Les variables de décision se retrouvent dans le tableau 5.12.

Tableau 5.11 Ensembles du problème.

Notation	Signification
P	Ensemble des périodes
J	Ensemble des quarts réguliers
JP	Ensemble des quarts à temps partiel
H	Ensemble des heures supplémentaires
M	Ensemble des pauses
Ξ	Ensemble des événements aléatoires

Tableau 5.12 Variables de décision du problème

Notation	Signification
S_j	Nombre d'employés assignés au quart $j \in J$
SP_j^u	Nombre d'employés assignés au quart à temps partiel $j \in JP$ pour la réalisation $u \in \Xi$
SH_h^u	Nombre de fois où l'heure supplémentaire $h \in H$ est attribuée pour la réalisation $u \in \Xi$
B_m^u	Nombre d'employés à temps plein utilisant la pause $m \in M$ pour la réalisation $u \in \Xi$
X_{jm}^u	Nombre d'employés assignés au quart j et utilisant la pause m pour la réalisation $u \in \Xi$
XH_{jh}^u	Nombre d'employés assignés au quart j et effectuant l'heure supplémentaire h pour la réalisation $u \in \Xi$
SC_p^u	Nombre d'employés manquants à la période $p \in P$ afin de répondre à la demande pour la réalisation $u \in \Xi$ (sous-couverture)

Finalement, les constantes apparaissant dans le programme stochastique sont expliquées dans le tableau 5.13.

Avec la notation définie ci-haut, la fonction objectif se formule comme suit :

$$\min \quad \sum_{j \in J} c_j S_j + \sum_{u \in \Xi} p_u \left[\sum_{j \in JP} cp_j SP_j^u + \sum_{h \in H} cs_h SH_h^u + \sum_{p \in P} csc \; SC_p^u \right] \tag{5.39}$$

Tableau 5.13 Constantes du problème.

Notation	Signification
c_j	Coût du quart $j \in J$
cp_j	Coût du quart à temps partiel $j \in JP$
cs_h	Coût du bloc d'heures supplémentaires $h \in H$
csc	Coût de la sous-couverture (identique pour toutes les périodes)
p_u	Probabilité de l'événement $u \in \Xi$
D_p^u	Demande à la période $p \in P$ pour la réalisation $u \in \Xi$
A_{jp}	Vaut 1 si le quart $j \in J$ couvre la période $p \in P$, 0 sinon
AP_{jp}	Vaut 1 si le quart à temps partiel $j \in JP$ couvre la période $p \in P$, 0 sinon
AK_{mp}	Vaut 1 si la pause $m \in M$ couvre la période $p \in P$, 0 sinon
AH_{hp}	Vaut 1 si le bloc d'heures supplémentaires $h \in H$ couvre la période $p \in P$, 0 sinon
Q_{jm}	Vaut 1 si la pause $m \in M$ est compatible avec le quart $j \in J$, 0 sinon
R_{jh}	Vaut 1 si le bloc d'heures supplémentaires $h \in H$ est compatible avec le quart $j \in J$, 0 sinon

et est soumise aux contraintes suivantes :

$$\text{s.c.} \quad \sum_{j\in J} A_{jp}S_j + \sum_{j\in JP} AP_{jp}SP_j^u - \sum_{m\in M} AK_{mp}B_m^u + \sum_{h\in H} AH_{hp}SH_h^u + SC_p^u \geq D_p^u \qquad \forall\ (u,p) \in (\Xi, P) \quad (5.40)$$

$$\sum_{m\in M} Q_{jm}X_{jm}^u - S_j = 0 \qquad \forall\ (u,j) \in (\Xi, J) \quad (5.41)$$

$$\sum_{j\in J} Q_{jm}X_{jm}^u - B_m^u = 0 \qquad \forall\ (u,m) \in (\Xi, M) \quad (5.42)$$

$$\sum_{h\in H} R_{jh}XH_{jh}^u \leq S_j \qquad \forall\ (u,j) \in (\Xi, J) \quad (5.43)$$

$$\sum_{j\in J} R_{jh}XH_{jh}^u - SH_h^u = 0 \qquad \forall\ (u,h)\ \in (\Xi, H) \quad (5.44)$$

$$S_j,\ SP_j^u,\ SH_h^u,\ B_m^u,\ X_{jm}^u,\ XH_{jh}^u \in \mathbb{Z}^+ \qquad \forall\ (u,j,h,m) \in (\Xi, J, H, M). \quad (5.45)$$

Le seul paramètre aléatoire de ce problème correspond à la demande $\mathbf{D} = (D_1, ..., D_{96})$. La génération de scénarios pour ce problème consiste donc à discrétiser la demande par un ensemble de vecteurs $\mathbf{D}^j = (D_1^j, ..., D_{96}^j)$, $j = 1, ..., K$ auxquels on associe des probabilités p_j.

Puisque la demande en employés ne prend que des valeurs entières positives, on cherche des scénarios $\mathbf{D}^j \in \mathbb{N}^{96}\ \forall\ j$. Nous résolvons ici le PS à partir d'un ensemble de données historiques provenant du secteur de la restauration, qui contient la demande en employés prévue par l'entreprise au moment de fabriquer les horaires ainsi que la demande réellement observée la journée même [3]. Les données ont été recueillies sur 4 semaines distancées dans l'année. La demande prend des valeurs entières et n'excède pas plus de 8 employés pour tous nos trois ensembles de données. Un exemple d'une réalisation de la demande en employés au cours d'une journée est représenté sur la figure 5.16.

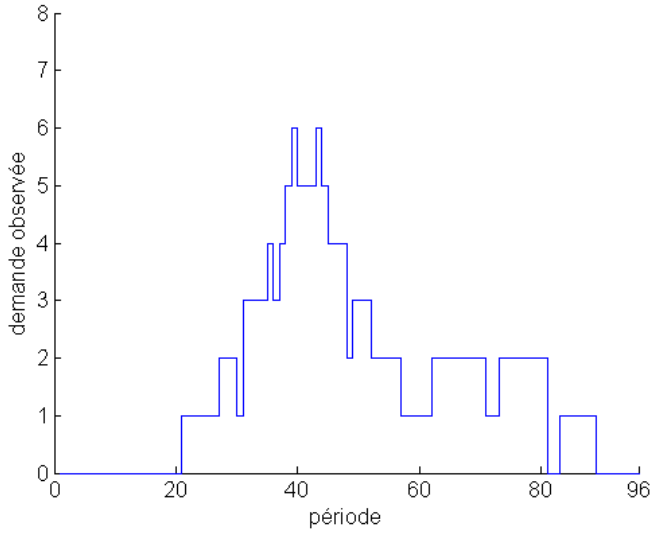

Figure 5.16 Exemple de réalisation de la demande en employés au cours d'une journée

Le programme stochastique étant posé, il faut réfléchir à quelle serait la meilleure méthode de génération de scénarios pour notre problème parmi celles présentées dans ce mémoire. Notons que la demande ne suit certainement pas la même loi pour tous les jours de la semaine. Par exemple, le nombre d'employés requis risque d'être différent la semaine et la fin de semaine. Il faudrait donc générer des scénarios pour chaque jour de la semaine individuellement. Or, on possède seulement 27 données [4] réparties sur 4 semaines de l'année, ce qui correspond à 4 données par jour de semaine, une quantité bien insuffisante pour la génération de scénarios. Pour remédier à ce problème, on profite du fait que l'on connaît également la demande en employés qui avait été prédite par les entreprises. La demande prévue tient compte entre

3. La demande en employés ne peut évidemment pas être mesurée directement. Elle est déduite par les entreprises à partir de facteurs tels que le nombre de transactions aux caisses.

4. Nous ne possédons pas 28 données puisque l'ensemble contenait un jour férié.

autres choses du jour de la semaine et de la période de l'année. Il est donc plausible que la différence

$$\theta_{np} = D_{np} - \widehat{D}_{np}, \quad p = 1, ..., 96 \ ; \ n = 1, ..., 28 \tag{5.46}$$

entre les demandes observée D_{np} et prévue $\widehat{D}_{np}$ de la p^{e} période suive la même loi indépendamment du jour n. On supposera donc que θ_{np} représente des observations d'une variable aléatoire θ_p. Ainsi, on commencera par générer les scénarios à partir des données θ_{np}. Ceux de la demande observée D_{np} seront ensuite obtenus en additionnant la demande prévue $\widehat{D}_{np}$ aux scénarios de θ_p.

Remarquons qu'il aurait aussi été possible de générer des scénarios à partir du rapport $D_{np}/\widehat{D}_{np}$. Cependant, ces rapports peuvent prendre des valeurs fractionnelles. Les scénarios de demande prévue ne seraient donc plus nécessairement entiers et il faudrait procéder à leur arrondissement. Cette relaxation linéaire est peut être envisageable lorsque la demande en employés est élevée, mais elle induit une grosse erreur d'approximation pour des valeurs plus faibles de D_{np}. L'écart relatif entre 70,6 et 71 employés, par exemple, est bien moins grand que celui entre 1,6 et 2 employés.

On vérifie maintenant s'il est possible d'estimer la distribution de θ_p, qui comprend l'estimation des distributions marginales, de leurs paramètres et de la matrice des corrélations. Puisqu'on ne possède que 27 données, il serait assez difficile d'estimer la distribution de notre vecteur aléatoire à 96 dimensions. On pourrait alors vérifier si la distribution de θ_p semble être la même pour certaines périodes. Le cas échéant, il serait possible d'estimer la loi de probabilité pour ces périodes à partir de la somme des données recueillies pour chacune d'entre elles. Or, la demande ne prend que des valeurs entières entre 0 et 8. Les distributions marginales θ_p doivent donc être représentées par des lois de probabilité entières. Bien que la QVAC puisse tirer avantage de la connaissance d'une distribution aléatoire continue, elle n'est malheureusement pas conçue pour quantifier des variables aléatoires discrètes. En fait, l'algorithme de Lloyd (discret) ne permet pas de générer des scénarios entiers non plus. Étant donné la faible quantité de données et la distribution entière de la demande, nous avons choisi de générer les scénarios à partir de l'algorithme d'échange des centres. Ce dernier garantit l'obtention de scénarios entiers, puisqu'il les sélectionne parmi les points de données.

L'algorithme d'échange des centres tel que présenté dans ce mémoire choisit les scénarios selon la norme L_1. Il serait également possible de choisir les centres en utilisant la distance L_2. Nous avons vu dans plusieurs cas que la qualité des solutions obtenues par les deux normes était très similaire. Nous avons cependant préféré utiliser la norme L_1 puisqu'il est plus pro-

bable pour la médiane $\nu_p^{1/2}$ d'une dimension p de prendre une valeur entière[5] que pour sa moyenne, ce qui simplifiera la procédure de génération de scénarios présentée ci-dessous.

L'algorithme d'échange des centres servira donc à partitionner notre ensemble de $N = 27$ données de dimension $D = 96$ pour former les scénarios. Nous avons vu que la minimisation de la distance de Wasserstein par l'algorithme d'échange des centres est plus difficile en dimension élevée, mais ne requiert pas plus de temps (voir tableau 5.8). Nous savons également que l'analyse par composantes principales peut être bénéfique en dimension élevée lorsqu'il y a présence de corrélations ou de grandes différences entre les variances marginales. Malheureusement, elle ne permet pas d'obtenir des scénarios entiers. Basés sur ces résultats, nous proposons l'heuristique suivante dont le principe est similaire à l'ACP, soit de réduire la dimension effective du problème tout en conservant une grande proportion des variances et covariances :

1. Sélectionner un ensemble de dimensions p dont la distance moyenne δ_p par rapport à la médiane est supérieure ou égale à une certaine valeur critique δ^*
2. Former des sous-ensembles à partir des dimensions qui sont les plus corrélées entre elles
3. Générer des scénarios «locaux» pour chaque sous-ensemble avec l'algorithme d'échange des centres
4. Former des scénarios «globaux» par combinaison des scénarios locaux

Nous expliquons maintenant plus en détail les étapes de l'heuristique à partir des données provenant du domaine de la restauration. Le magasin considéré est ouvert pour 64 périodes de 15 minutes au cours de la journée (sur une possibilité de 96). Notre problème est donc réduit à 64 dimensions, ce qui reste tout de même beaucoup trop pour obtenir une génération de scénarios de qualité. On procède donc à une nouvelle réduction de la dimension en première étape de l'heuristique, où l'on conserve seulement celle dont la distance moyenne par rapport à la médiane :

$$\delta_p = \frac{1}{N} \sum_{n=1}^{N} |\theta_{np} - \nu_p^{1/2}| \quad , \; p = 1, ..., P \tag{5.47}$$

est supérieure ou égale à une valeur critique δ^*. Il est ainsi possible d'éliminer (resp. conserver) autant de dimensions que l'on désire en augmentant (resp. diminuant) la valeur de δ^*. La justification de cette procédure provient du fait que si la valeur de δ_p est petite pour une dimension donnée, alors la demande de cette période sera bien représentée par sa valeur médiane. La figure 5.17 représente la distance moyenne par rapport à la médiane de

5. En fait, la médiane sera toujours entière dans notre cas puisque nous avons un nombre impair de données (27).

chaque dimension, où la droite horizontale correspond à la valeur critique choisie pour cet ensemble de données. On observe que les périodes en milieu de journée possèdent une plus grande variance de la demande, ce qui est très plausible, puisqu'elles sont normalement les plus achalandées.

Sur les 64 périodes pendant lesquelles le magasin est ouvert, on aimerait en conserver entre 30 et 40 pour la génération de scénarios. De cette manière, il sera plus facile de les répartir en sous-ensembles de taille raisonnable en 2^e étape. On trouve alors que le seuil devrait se situer entre 19/27 et 22/27. On possède respectivement 4, 1, 3 et 8 périodes dont la distance moyenne par rapport à la médiane se trouve entre 19/27 et 22/27. Nous avons donc choisi d'établir le seuil à $\delta^* = 20/27$, qui correspond à la distance par rapport à la médiane d'une seule dimension et permet d'obtenir un plus grand écart entre les dimensions considérées et rejetées. Avec ce choix pour δ^*, 30 dimensions sont éliminées du problème de génération de scénarios et seront représentées par leur valeur médiane.

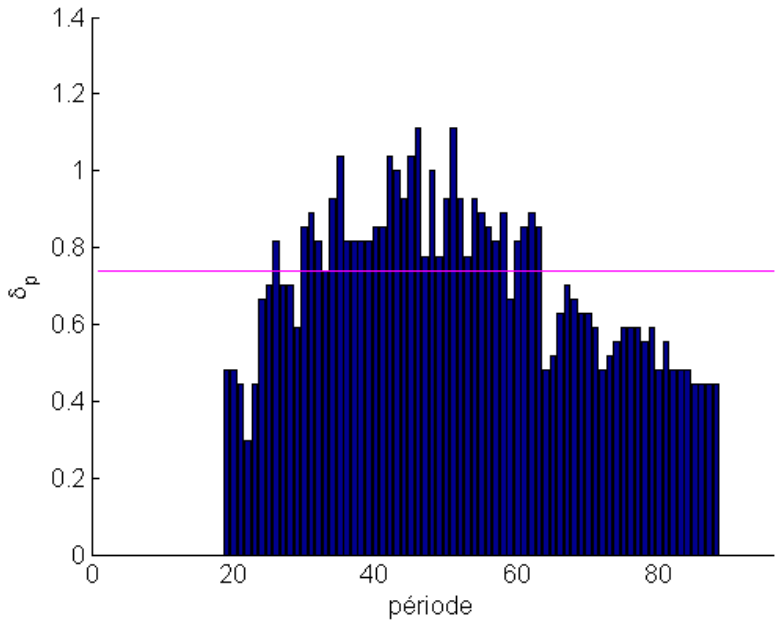

Figure 5.17 Distances moyennes des demandes de chaque période par rapport à leur médiane

Il nous reste 34 périodes pour lesquelles nous devons générer les scénarios marginaux, ce qui représente encore un grand nombre de dimensions. Nous les avons donc divisés en sous-ensembles Φ_m, $m = 1, ..., M$ de périodes possédant de fortes corrélations. Pour ce faire, on

commence par regrouper les dimensions dont l'estimateur de la corrélation défini par

$$r_{pq} = \frac{\sum_{n=1}^{N}(\theta_{np} - \overline{\theta}_p)(\theta_{nq} - \overline{\theta}_q)}{\sqrt{\sum_{n=1}^{N}(\theta_{np} - \overline{\theta}_p)^2 \sum_{n=1}^{N}(\theta_{nq} - \overline{\theta}_q)^2}} \tag{5.48}$$

(où $\overline{\theta}_p$ correspond à la moyenne arithmétique de θ_p pour la période p) est supérieur à une certaine valeur critique r^*. Si deux dimensions possèdent une corrélation supérieure à r^*, alors elles doivent se retrouver dans le même sous-ensemble. Puisque deux périodes peu espacées dans le temps ont plus de chance d'être corrélées, on associe celles qui n'ont aucune corrélation estimée supérieure à r^* au sous-ensemble Φ_m avec lequel l'écart de temps est le plus petit. Si une période est située exactement entre 2 sous-ensembles, alors nous avons arbitrairement choisi de l'associer au sous-ensemble le plus tôt. Cette procédure est schématisée sur la figure 5.18, où nous avons choisi une corrélation seuil $r^* = 0,6$. Cette décision repose sur le fait qu'il ne nous semble pas essentiel que les scénarios des périodes dont la corrélation est inférieure à ce seuil soient générés ensemble, tandis que celles qui ont une corrélation supérieure à 0,6 doivent clairement être considérées simultanément.

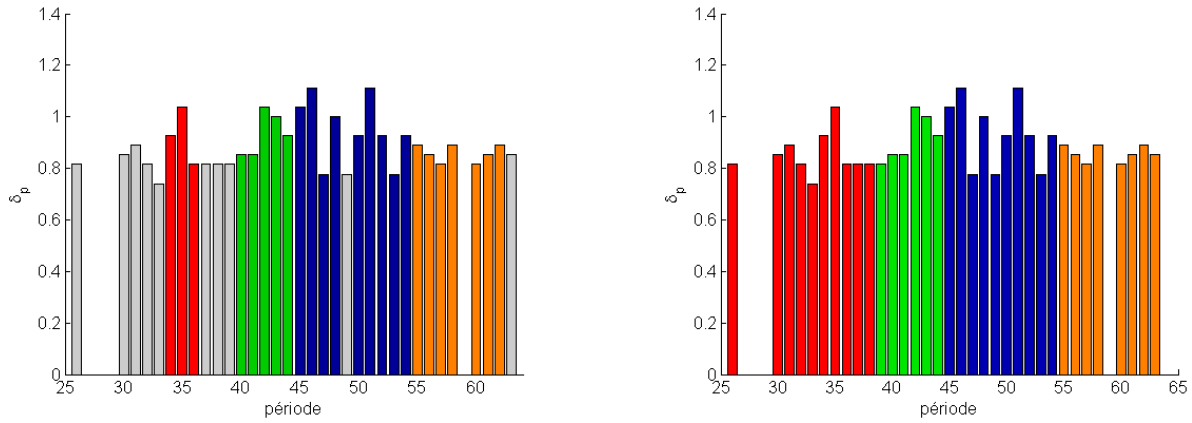

Figure 5.18 Sous-ensembles de dimensions fortement corrélées (couleur) (figure de gauche). Association des dimensions faiblement corrélées (gris) aux sous-ensembles les plus près (figure de droite).

Les 34 dimensions restantes sont donc divisées en $M = 4$ sous-ensembles Φ_m, $m = 1, ..., 4$ de 10, 6, 10 et 8 périodes respectivement (énumérés en ordre croissant dans le temps). On constate qu'il n'y a pas de «croisements» entre les dimensions appartenant à chaque sous-ensemble. C'est-à-dire que si deux périodes font partie de Φ_m pour un m quelconque, alors toutes celles qui se trouvent entre les deux premières appartiennent également à Φ_m. Ce

résultat est logique puisqu'intuitivement, il est clair que deux périodes successives ont plus de chance d'être corrélées. La procédure choisie pour associer les dimensions restantes aux sous-ensembles initiaux permet également d'éviter les croisements possibles entre les périodes.

Ayant réduit les 64 dimensions de notre problème initial, nous pouvons maintenant procéder à la génération de scénarios des vecteurs $\boldsymbol{\theta}_m$, $m = 1, ..., M$ formés à partir des distributions marginales des dimensions de Φ_m, possédant tous moins de 11 périodes. Puisque les périodes semblent toutes avoir environ la même influence sur le programme stochastique (5.39) pour chaque groupe Φ_m, il n'est pas nécessaire d'appliquer la transformation présentée à la proposition 3.4. Tel que mentionné précédemment, nous avons choisi d'utiliser l'algorithme d'échange des centres, puisqu'il permet d'obtenir des scénarios entiers de $\boldsymbol{\theta}_m$ et de la demande par conséquent. Les scénarios obtenus à cette étape seront dits «locaux», puisqu'il représente seulement une partie des dimensions de la demande journalière $\mathbf{D} = (D_1, ..., D_{96})$. Nous avons choisi la quantité de scénarios locaux K_m proportionnellement au nombre de dimensions des sous-ensembles Φ_m. Les scénarios «globaux» sont formés par combinaison des scénarios locaux, d'une manière analogue à ce qui avait été fait pour l'heuristique de quadrillage. Le programme stochastique (5.39) peut être résolu avec $K = 500$ scénarios (globaux) en un temps raisonnable à l'aide de l'algorithme développé par Pacqueau (2011). Celui-ci correspond à une heuristique basée sur la méthode *L-shaped*, où l'on résout d'abord (5.39) en appliquant une relaxation linéaire aux sous-problèmes (il y en a un par scénario), puis en fixant les variables dont la partie fractionnaire est supérieure à 0,8. On résout la relaxation linéaire à nouveau et on répète le processus jusqu'à ce que toutes les variables soient fixées. On obtient finalement une solution en procédant à la résolution en valeur entière du problème maître.

Ainsi, en choisissant $K_m = |\Phi_m|/2$, $m = 1, ..., 4$, où $|\Phi_m|$ est le cardinal de Φ_m, on obtient respectivement 5, 3, 5 et 4 scénarios locaux de $\boldsymbol{\theta}_m$ pour un total de $K = 5 \times 3 \times 5 \times 4 = 300$ scénarios globaux de $\boldsymbol{\theta} = (\theta_1, ..., \theta_{96})$, puisque les autres dimensions sont représentées par leur médiane. Les scénarios de la demande en employés pour une journée n sont finalement obtenus en additionnant les scénarios $\boldsymbol{\theta}$ et la demande prévue $\widehat{D}_{np}$ (voir l'équation (5.46)).

Notons que l'heuristique ci-dessus génère une quantité de scénarios supérieure aux données fournies. Si l'on appliquait simplement l'algorithme d'échange des centres à partir des données, il ne serait alors pas possible de générer plus de $N = 27$ scénarios pour chaque journée. Un avantage de la méthode proposée est qu'elle permet une optimisation plus robuste en formant des scénarios qui n'ont pas été observés directement, mais dont l'existence est plausible.

Le programme stochastique (5.39) a été résolu pour chaque journée de données avec les différentes instances du vecteur aléatoire de la demande présentées dans le tableau 5.14. Les solutions optimales x^* sont obtenues par l'instance OBS où le vecteur «aléatoire» correspond à la demande observée (avec probabilité 1). Elles correspondent aux décisions que l'on prendrait s'il était possible de prédire le futur pour chaque journée. Pareillement, le PS [6] est résolu à partir de la demande qui avait été prévue par l'entreprise pour l'instance PRE. Le seul problème d'optimisation stochastique traité (OS) est celui où les scénarios sont générés par la procédure décrite ci-haut tandis que l'optimisation déterministe (OD) ne considère que la somme du vecteur médian de θ_{np} et de la demande prévue

$$D_{np}^{(OD)} = \nu_p^{1/2} + \widehat{D}_{np} \qquad \forall\ n \in \{1, ..., N\},\ p \in \{1, ..., 96\}\ . \tag{5.49}$$

Tableau 5.14 Instances de la demande

Notation	Instance
OBS	Demande observée
PRE	Demande prévue
OS	Optimisation stochastique à partir de 300 scénarios
OD	Optimisation déterministe (scénario médian)

La résolution du PS avec la demande observée nous permet de trouver le coût optimal $Z(x^*)$ de notre problème. Pour évaluer la qualité des autres scénarios, on mesure le coût réel de leurs solutions sur le programme stochastique OBS (voir équation (3.7)). Les coûts réels des instances OBS, PRE et OS sont présentés à la figure 5.19 pour les 27 journées de données.

On constate que l'optimisation stochastique permet des économies modérées sur les dépenses liées à la confection d'horaire pour la majorité des journées. Les moyennes et variances échantillonnales des coûts réels sur les 27 journées sont présentées dans le tableau 5.15.

Tableau 5.15 Moyennes et variances échantillonnales des coûts réels

	OBS	OS	PRE
$E[Z]$	51,342	54,372	55,105
$Var[Z]$	-	36,184	44,275

6. En fait, il ne s'agit plus réellement d'un programme *stochastique* puisqu'il n'y a qu'une seule représentation de la variable aléatoire

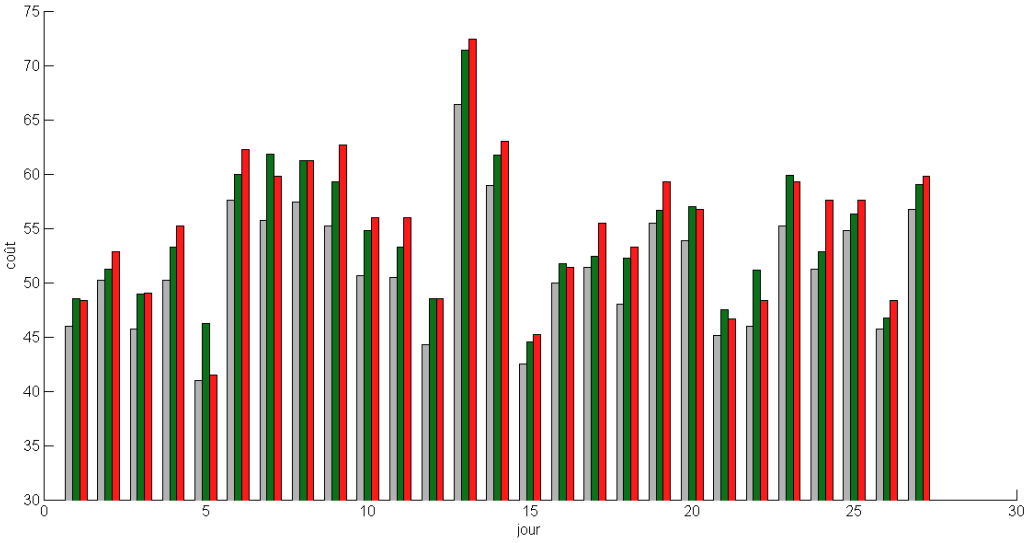

Figure 5.19 Coûts réels de solutions pour les instances OBS (gris), OS (vert) et PRE (rouge)

Si l'on était capable de prédire le futur, alors le coût quotidien de la confection d'horaires pour l'entreprise étudiée serait de 51,342 unités. On constate que l'optimisation stochastique permet effectivement de réduire les dépenses journalières moyennes de la confection d'horaires. L'optimisation basée sur la demande prévue engendre des dépenses moyennes de 55,105 unités par jour, tandis que l'OS conduit à des coûts moyens de 54,372 unités, ce qui représente des économies de 1,33 %. Ce gain peut paraître faible, mais présente tout de même un intérêt substantiel considérant que la confection d'horaire correspond souvent à la dépense principale des entreprises. De plus, la variance des coûts de l'OS est plus faible que celle de l'optimisation basée sur la demande prévue. L'optimisation stochastique permet donc non seulement d'abaisser les dépenses en moyenne, mais réduit également les risques économiques qui y sont associés.

En toute justice, il faut noter que le résultat ci-dessus est peut-être légèrement biaisé. Idéalement, il aurait fallu diviser les données en deux ensembles : le premier pour générer les scénarios et l'autre pour les tester. Malheureusement, puisque l'on ne possédait pas une grande quantité de données, nous avons été contraints de les utiliser à la fois pour la génération de scénarios et pour leur évaluation. Il se peut donc que les profits observés ne soient pas dus à l'optimisation stochastique, mais plutôt au fait que l'on ait tiré avantage de la connaissance de la demande *a posteriori*.

Nous avons donc également évalué le coût réel des solutions obtenues en ne considérant que le vecteur médian $D_{np}^{(OD)}$ de la demande donné par la formule (5.49). Nous avons mesuré un coût journalier moyen de 54,635 unités et une variance échantillonnale de 38,703 pour cette instance de la demande. Bien qu'ils soient assez faibles, l'optimisation stochastique permet de réaliser des gains de 0,48 % et possède une variance légèrement inférieure à l'instance OD. Bien entendu, il aurait été plus satisfaisant d'observer des profits plus importants. Ces gains sont probablement tout de même suffisants pour justifier l'utilisation de l'OS dans certains cas, mais tout dépend du budget de l'entreprise accordé à la fabrication d'horaires. Il faut aussi noter que l'écart relatif entre les dépenses minimales (OBS) et ceux de l'OD n'est que de 6,10 %. Ainsi, même si l'on était capable de prévoir l'avenir, les gains observés (aussi appelés valeur de l'information parfaite) ne seraient pas énormes pour ce problème. De plus, Pacqueau (2011) a constaté une dépendance entre les gains de l'OS et le coefficient de variation moyen donné par

$$\gamma = \frac{1}{NP} \sum_{n=1}^{N} \sum_{p=1}^{P} \left(\frac{1}{D_{np}^{(OD)}} \sqrt{\frac{1}{K} \sum_{j=1}^{K} (D_{njp}^{(OS)} - D_{np}^{(OD)})} \right) \ , \tag{5.50}$$

où $D_{njp}^{(OS)}$ représente la demande de la période p du jour n pour le scénario j de OS. Lorsque $D_{np}^{(OD)}$ est nul, on pose les termes à l'intérieur des sommations égaux à 0. Pour $\gamma < 0,49$, les gains de l'OS étaient inférieurs à 0,7 % pour toutes les instances étudiées avec 500 scénarios. Dans notre cas, on calcule un coefficient de variation moyen $\gamma = 0,168$, ce qui est très faible et seulement 300 scénarios sont utilisés. Il est donc tout à fait naturel que la valeur de la solution stochastique soit assez faible. Néanmoins, l'optimisation de la confection d'horaires nous a permis de constater des gains à partir de données historiques sur un problème concret, malgré la dimension élevée et le faible coefficient de variation moyen.

CHAPITRE 6

CONCLUSION

6.1 Synthèse des travaux

Le travail présenté dans ce mémoire se résume principalement aux aspects suivants :

1. La présentation détaillée de résultats théoriques sur la quantification optimale justifiant les méthodes proposées
2. Le développement d'algorithmes de partitionnement de données pour les distributions connues et inconnues
3. L'analyse de certaines caractéristiques, telles que le nombre de scénarios et la dimension du problème, sur la génération de scénarios
4. L'évaluation de nos méthodes sur le problème analytique du vendeur de journaux et un problème concret de confection d'horaires

La quantification optimale nous a permis de trouver une borne supérieure sur l'erreur de discrétisation du programme stochastique. Celle-ci dépend notamment de la distance de Wasserstein, une métrique probabiliste qui mesure la distance moyenne entre une réalisation du vecteur aléatoire et le scénario le plus près. Ce résultat est à la base de la méthode de génération de scénarios proposée, qui cherche à minimiser la distance de Wasserstein. Nous avons démontré de façon détaillée certains résultats par rapport aux métriques probabilistes qui nous ont permis de faire le lien entre la minimisation de la distance de Wasserstein et les problèmes de partitionnement de données.

Il peut parfois être difficile d'estimer avec confiance la distribution d'un vecteur aléatoire. Nous avons donc proposé des algorithmes de partitionnement de données adaptés à la connaissance de la loi probabiliste. Lorsque la distribution est connue, la méthode de Lloyd converge vers le minimum global du problème de partitionnement en une dimension. Pour le cas multidimensionnel, nous avons utilisé la quantification vectorielle par apprentissage compétitif, qui permet d'obtenir des valeurs satisfaisantes de la distance de Wasserstein. Des algorithmes de QVAC existent déjà pour la norme L_2, mais nous avons trouvé peu d'informations pour les problèmes avec la norme en valeur absolue. Nous avons donc ajusté les paramètres de la QVAC avec la norme L_1 pour nous permettre d'évaluer l'erreur de discrétisation de l'optimisation stochastique. Lorsque la distribution est inconnue, nous avons présenté les algorithmes

d'échange des centres et de Lloyd, qui génèrent les scénarios directement à partir des données.

Nous avons ensuite analysé les effets du nombre de scénarios, de la dimension du problème, de la matrice des covariances et de la norme sur la génération de scénarios. Les résultats théoriques sur l'erreur de discrétisation ont été obtenus avec la norme L_1. Cependant, on a trouvé qu'il était également possible de générer les scénarios à partir de la distance euclidienne. Cette dernière ne permet pas d'évaluer la borne supérieure sur l'erreur de discrétisation, mais nos résultats démontrent que cette évaluation offre peu d'intérêt puisque l'erreur obtenue est généralement bien en deçà de la borne supérieure. Le choix de la norme dépend du programme stochastique et de la représentation désirés des événements extrêmes. Avec une quantité de scénarios suffisante, la distance de Wasserstein obtenue est toujours inférieure à celle de l'échantillonnage pur. On trouve cependant que la génération de scénarios est particulièrement difficile lorsque la dimension est élevée. Nous avons donc proposé trois méthodes de réduction de la dimension : une heuristique de quadrillage, l'utilisation des copules et l'analyse par composantes principales. Seule la dernière des techniques proposées conduit à de bons résultats et permet de réduire la distance de Wasserstein obtenue par la QVAC d'environ 20 % avec un choix approprié du nombre de composantes principales en 25 dimensions.

Nous avons vérifié à l'aide du problème du vendeur de journaux si la minimisation de la distance de Wasserstein conduit effectivement vers une erreur de discrétisation minimale pour le programme stochastique. Même si une distance de Wasserstein inférieure ne garantit pas l'obtention d'une meilleure solution, on observe effectivement une tendance des coûts du programme stochastique à décroître lorsqu'elle est faible. En une dimension, l'algorithme de Lloyd donne une erreur de discrétisation plus de deux fois plus petite que celle de l'échantillonnage pur et environ 350 fois inférieure à celle de l'optimisation déterministe. Pour le cas multidimensionnel, une plus grande quantité de scénarios doit cependant être utilisée pour assurer l'obtention de solutions de qualité.

Nous avons finalement évalué l'algorithme d'échange des centres sur un problème concret lié à la confection d'horaires de personnel. Ce programme stochastique représentait un défi particulier puisqu'il comportait 64 dimensions et un faible coefficient de variation moyen. Nous avons tout de même réussi à proposer une heuristique qui réduit la dimension effective du problème en conservant une grande proportion des variances marginales et ainsi obtenir des gains de 0,5 % avec seulement 300 scénarios par rapport à l'optimisation déterministe. L'OS peut donc être profitable chez certaines entreprises pour lesquelles le budget lié à

la confection d'horaire est assez grand. En général, l'optimisation stochastique sera plus avantageuse lorsque le problème comporte moins de dimensions ou des variances plus élevées.

6.2 Limitations de la solution proposée

On trouve généralement que l'erreur de discrétisation est beaucoup plus petite que la borne supérieure trouvée théoriquement par la quantification optimale. Ainsi, notre méthode ne permet pas d'obtenir une évaluation de l'erreur qui possède un réel intérêt.

Lorsque la distribution est connue, la QVAC converge rapidement vers une discrétisation du vecteur aléatoire minimisant la distance de Wasserstein. Le temps de résolution de l'algorithme d'échange des centres est toutefois bien plus long. Il n'est donc pas possible de quantifier un vecteur aléatoire par cette méthode lorsque la quantité de données et le nombre de scénarios recherchés sont trop élevés. L'algorithme de Lloyd converge rapidement vers des minimums locaux, mais ne permet pas de trouver des scénarios discrets, tels que la demande en employés pour le problème de confection d'horaires étudié. Dans ces cas, il faudrait considérer d'autres techniques de partitionnement de données dont la convergence est plus rapide.

Lorsque la dimension est élevée, même si l'analyse par composantes principales permet parfois de réduire les coûts liés à l'optimisation stochastique, il peut être nécessaire de considérer une très grande quantité de scénarios afin d'obtenir des solutions de qualité. Or, certains programmes stochastiques plus complexes ne peuvent inclure un grand nombre de scénarios en maintenant un temps de convergence acceptable. L'optimisation stochastique a donc de meilleures chances d'être profitable pour les PS plus simples.

Nous avons eu la chance de travailler sur un problème concret à partir de données historiques. Les résultats sont assez encourageants, mais certaines méthodes telles que la QVAC n'ont pu être évaluées puisque le vecteur aléatoire ne prenait que des valeurs entières. Les méthodes proposées devraient donc être testées sur un plus grand nombre d'applications pour réellement apprécier leur utilité.

6.3 Améliorations futures

Parmi les aspects qui n'ont pas été étudiés dans ce mémoire, un des plus intéressants serait sans doute d'établir des critères permettant de discerner les ensembles de données pour lesquels il est préférable d'estimer le vecteur aléatoire de ceux pour lesquels il vaut mieux choisir les scénarios directement à partir des données. Par exemple, quel intervalle de confiance sur

l'estimation des paramètres du vecteur aléatoire est nécessaire afin que la QVAC soit avantageuse par rapport à l'algorithme d'échange des centres ? Un plus grand nombre de données avantage-t-il une méthode plus que l'autre ?

Il existe évidemment plusieurs méthodes de génération de scénarios. Nous n'en avons essentiellement considéré qu'une seule dans ce mémoire (avec certaines variantes). Il pourrait donc être intéressant de confronter la quantification optimale à d'autres méthodes de génération de scénarios. Existe-t-il des instances pour lesquelles certaines méthodes sont plus efficaces que d'autres ? Le cas échéant, est-il possible d'extraire les caractéristiques du problème afin de sélectionner la méthode la plus performante ?

Il existe également une énorme quantité d'algorithmes des k-moyennes et k-médianes qui pourraient être utilisés afin de minimiser la distance de Wasserstein et qui méritent d'être explorés en fonction de la quantité de données disponibles, du nombre de scénarios requis et de la dimension du problème.

Finalement, comme il été dit précédemment, certaines des méthodes présentées dans ce mémoire n'ont pu être testées sur des applications concrètes. Il serait donc nécessaire de les évaluer sur des problèmes tangibles afin de réellement vérifier leur utilité.

RÉFÉRENCES

BENDERS, J. F. (1962). Partitioning procedures for solving mixed-variables programming problems. *Numerische mathematik*, 4, 238–252.

BIRGE, J. et LOUVEAUX, F. (2011). *Introduction to Stochastic Programming*. Springer Series in Operations Research and Financial Engineering. Springer New York.

CARDOT, H., CENAC, P. et ZITT, P.-A. (2013). Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19, 18–43.

CARINO, D., KENT, T., STACY, C., SYLVANUS, M., TURNER, A., WATANABE, K. et ZIEMBA, W. (1994). The russell-yasuda kasai model : an asset/liability model for a japanese insurance company using multistage stochastic programming. *Interfaces*, 24, 29 – 49.

CHAN, Z., COLLINS, L. et KASABOV, N. (2006). An efficient greedy k-means algorithm for global gene trajectory clustering. *Expert Systems with Applications*, 30, 137 – 41.

CHARIKAR, M., GUHA, S., TARDOS, E. et SHMOYS, D. B. (2003). A constant-factor approximation algorithm for the k-median problem. vol. 65, 129 – 149.

CHEN, K. (2009). On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM Journal on Computing*, 39, 923 – 47.

DU, Q., EMELIANENKO, M. et JU, L. (2006). Convergence of the lloyd algorithm for computing centroidal voronoi tessellations. *SIAM Journal on Numerical Analysis*, 44, 102–119.

DUDLEY, R. M. (1976). *Probabilities and metrics : Convergence of laws on metric spaces, with a view to statistical testing*, vol. 45. Aarhus Universitet, Matematisk Institut.

ERMOLIEV, Y. (1983). Stochastic quasigradient methods and their application to system optimization. *Stochastics : An International Journal of Probability and Stochastic Processes*, 9, 1–36.

EVERITT, B. S., LANDAU, S., LEESE, M. et STAHL, D. (2011). *An Introduction to Classification and Clustering*. John Wiley & Sons, Inc.

GRAHAM, R. L., KNUTH, D. E. et PATASHNIK, O. (1994). *Concrete Mathematics : A Foundation for Computer Science*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, seconde édition.

GULPINAR, N., RUSTEM, B. et SETTERGREN, R. (2004). Simulation and optimization approaches to scenario tree generation. *Journal of Economic Dynamics & Control*, 28, 1291–1315.

HEYDE, C. (2010). On a property of the lognormal distribution. R. Maller, I. Basawa, P. Hall et E. Seneta, éditeurs, *Selected Works of C.C. Heyde*, Springer New York, Selected Works in Probability and Statistics. 16–18.

HIGLE, J. et SEN, S. (1991). Stochastic decompositino - an algorithm for 2-stage linear programs with recourse. *Mathematics of Operations Research*, 16, 650–669.

HOCHREITER, R. et PFLUG, G. (2007). Financial scenario generation for stochastic multi-stage decision processes as facility location problems. *Annals of Operations Research*, 152, 257 – 72.

HUBER, P. J. (2005). *The Weak Topology and its Metrization*, John Wiley & Sons, Inc. 20–42.

HYLAND, K., KAUT, M. et WALLACE, S. W. (2003). A heuristic for moment-matching scenario generation. *Computational Optimization and Applications*, 24, 169 – 185.

HYLAND, K. et WALLACE, S. W. (2001). Generating scenario trees for multistage decision problems. *Management Science*, 47, 295 – 307.

INDYK, P. (1999). Sublinear time algorithms for metric space problems. *Conference Proceedings of the Annual ACM Symposium on Theory of Computing*, 428 – 434.

INFANGER, G. (1992). Monte carlo (importance) sampling within a benders decomposition algorithm for stochastic linear programs. *Annals of Operations Research*, 39, 69–95.

JAIN, K. et VAZIRANI, V. V. (2001). Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and lagrangian relaxation. *Journal of the ACM*, 48, 274 – 296.

JAMSHIDIAN, F. et ZHU, Y. (1996). Scenario simulation : Theory and methodology. *Finance and Stochastics*, 1, 43–67.

JOLLIFFE, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer, seconde édition.

KANTOROVICH, L. V. et RUBINSTEIN, G. S. (1958). On a space of completely additive functions. *Vestnik Leningrad. Univ.*, 13, 52–59.

KAUT, M. (2013). A copula-based heuristic for scenario generation. *Computational Management Science*, 25.

KAUT, M. et WALLACE, S. W. (2003). Evaluation of scenario-generation methods for stochastic programming. *World Wide Web, Stochastic Programming E-Print Series*. 14–2003.

KAUT, M. et WALLACE, S. W. (2011). Shape-based scenario generation using copulas. *Computational Management Science*, 8, 181–199.

KEEFER, D. (1994). Certainty equivalents for three-point discrete-distribution approximations. *Management Science*, 40, 760 – 73.

KORUPOLU, M., PLAXTON, C. et RAJARAMAN, R. (2000). Analysis of a local search heuristic for facility location problems. *Journal of Algorithms*, 37, 146–188.

KOUWENBERG, R. (2001). Scenario generation and stochastic programming models for asset liability management. *European Journal of Operational Research*, 134, 279 – 292.

LATORRE, J. M., CERISOLA, S. et RAMOS, A. (2007). Clustering algorithms for scenario tree generation : Application to natural hydro inflows. *European Journal of Operational Research*, 181, 1339–1353.

LEGRAIN, A. (2012). *Generation de scénarios pour la demande en personnels durant plusieurs periodes*. Mémoire de maîtrise, École Polytechnique de Montréal.

LEMIEUX, C. (2009). Variance reduction techniques. *Monte Carlo and Quasi-Monte Carlo Sampling*, Springer New York, Springer Series in Statistics. 1–52.

LEVIN, V. L. et MILYUTIN, A. A. (1979). The problem of mass transfer with a discontinuous cost function and a mass statement of the duality problem for convex extremal problems. *Russian Mathematical Surveys*, 34, 1–78.

LIKAS, A., VLASSIS, N. et J. VERBEEK, J. (2002). The global k-means clustering algorithm. *Pattern Recognition*, 36, 451 – 461.

LLOYD, S. (1982). Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28, 129–137.

MILLER, A.C., I. et RICE, T. (1983). Discrete approximations of probability distributions. *Management Science*, 29, 352 – 62.

MOULINES, E. et BACH, F. R. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira et K. Weinberger, éditeurs, *Advances in Neural Information Processing Systems 24*. 451–459.

PACQUEAU, R. (2011). *Optimisation stochastique d'horaires de personnel*. Mémoire de maîtrise, École Polytechnique de Montréal.

PAGES, G. et PRINTEMS, J. (2003). Optimal quadratic quantization for numerics : the gaussian case. *Monte Carlo Methods and Applications*, 9, 135 – 65.

PEARSON, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50, 157–175.

PFLUG, G. (2001). Scenario tree generation for multiperiod financial optimization by optimal discretization. *Mathematical Programming*, 89, 251–271.

RACHEV, S., KLEBANOV, L., STOYANOV, S. et FABOZZI, F. (2013). Monge–kantorovich mass transference problem, minimal distances and minimal norms. *The Methods of Distances in the Theory of Probability and Statistics*, Springer New York. 109–143.

SKLAR, A. (1996). *Random variables, distribution functions, and copulas—a personal look backward and forward*, Institute of Mathematical Statistics, Hayward, CA, vol. Volume 28 de *Lecture Notes–Monograph Series*. 1–14.

SUTIENE, K. et PRANEVICIUS, H. (2007). Scenario generation employing copulas. *World Congress on Engineering*. Hong Kong, China, 777 – 84.

TOPALOGLOU, N., VLADIMIROU, H. et ZENIOS, S. (2002). CVaR models with selective hedging for international asset allocation. *Journal of Banking & Finance*, 26, 1535–1561.

VALLENDER, S. (1974). Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18, 784–786.

VAN SLYKE, R. M. et WETS, R. (1969). L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal on Applied Mathematics*, 17, 638–663.

VELMURUGAN, T. et SANTHANAM, T. (2010). Computational complexity between k-means and k-medoids clustering algorithms for normal and uniform distributions of data points. *Journal of Computer Sciences*, 6, 363 – 8.

WASSERSTEIN, L. (1969). Markov processes on a countable product space, describing large systems of automata (in russian). *Problemy Peredachi Infomatsii*, 5, 64–73.