Cytometry data modeling and unsupervised classification in moderately high dimensions under the independence structure assumption

Louis Pujol

université PARIS-SACLAY

Cytometry data modeling and unsupervised classification in moderately high dimensions under the independence structure assumption

Modélisation des données de cytométrie et classification non supervisée en dimension modérée sous l'hypothèse de structure d'indépendance

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 574, mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées
Graduate School : Mathématiques, Référent : Faculté des sciences d'Orsay

Thèse préparée dans le Laboratoire de Mathématiques d'Orsay (Université Paris-Saclay, CNRS), sous la direction de Pascal MASSART, professeur, et le co-encadrement de Marc GLISSE, chargé de recherche

Thèse soutenue à Paris-Saclay, le 1er décembre 2022, par

Louis PUJOL

Composition du jury
Membres du jury avec voix délibérative

Frédéric CHAZAL Directeur de recherche, INRIA Saclay	Président
Charles BOUVEYRON Professeur, Université Côte d'Azur	Rapporteur & Examinateur
Clémentine PRIEUR Professeure, Université Grenoble Alpes	Rapporteure & Examinatrice
Aurélie FISCHER Maîtresse de conférences, Université de Paris	Examinatrice

THESE DE DOCTORAT
NNT : 2022UPASM032

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre : Modélisation des données de cytométrie et classification non supervisée en dimension modérée sous l'hypothèse de structure d'indépendance.

Mots clés : Classification non supervisée,Estimation de densité,Analyse de données de cytométrie

Résumé : La cytométrie est une technique permettant de mesurer la présence de certaines protéines dans un échantillon cellulaire à l'échelle de la cellule individuelle. L'objectif est d'identifier des populations. Des approches manuelles reposants sur une expertise métier sont aujourd'hui majoritairement utilisées. Un enjeu du domaine est l'automatisation de cette tâche. Dans cette thèse, nous présentons une approche originale de classification non supervisée adaptée aux données de cytométrie. Elle repose sur l'enchaînement de deux étapes : une étape d'estimation de densité et une étape de classification non supervisée déterministe via l'algorithme ToMATo (Chazal, Guibas, Oudot, Skraba). L'étape d'estimation de densité est réalisée en prenant en compte une éventuelle hypothèse de structure d'indépendance dans les variables d'entrée. Ce modèle, introduit par Lepski et Rebelles, revient à supposer que l'on puisse séparer les variables en blocs indépendants. Nous montrons les bonnes performances de notre méthode sur des données de cytométrie en nous comparant à des études comparatives précédemment publiées, en particulier celle de Weber et Robinson. Nous présentons l'algorithme d'estimation de densité ISDE (Independence Structure Density Estimation) qui permet d'estimer une densité reposant sur une structure d'indépendance avec un temps de calcul raisonnable sur les tailles de données rencontrées en cytométrie. Nous montrons par un contrôle en grande probabilité du risque de Kullback-Leibler de l'estimateur obtenu que cette approche permet de réduire l'impact du fléau de la dimension. Enfin, nous montrons la pertinence du modèle de structure d'indépendance sur les données de cytométrie par une étude empirique de la qualité de l'estimation de densité obtenue par ISDE.

Title : Cytometry data modeling and unsupervised classification in moderately high dimensions under the independence structure assumption

Keywords : Clustering, Density estimation, Cytometry data analysis

Abstract : Cytometry is a technique for measuring the presence of certain proteins in a cell sample at the level of the individual cell. The objective is to identify populations. Manual approaches based on professional expertise are mostly used today. A challenge in this field is the automation of this task. In this thesis, we present an original unsupervised classification approach adapted to cytometry data. It is based on a sequence of two steps : a density estimation step and a deterministic unsupervised classification step via the ToMATo algorithm (Chazal, Guibas, Oudot, Skraba). The density estimation step is performed by taking into account a possible hypothesis of independence structure in the input variables. This model, introduced by Lepski and Rebelles, amounts to assuming that we can separate the variables into independent blocks. We show the good performance of our method on cytometry data by comparing with previously published comparative studies, in particular that of Weber and Robinson. We present the ISDE (Independence Structure Density Estimation) algorithm allowing to estimate a density based on an independence structure with a reasonable computation time on the sizes of data encountered in cytometry. We show by a high probability control of the Kullback-Leibler risk of the obtained estimator that this approach allows to reduce the impact of the curse of dimensionality. Finally, we show the relevance of the independence structure model on cytometry data by an empirical study of the quality of the density estimate obtained by ISDE.

Remerciements

Je remercie les membres du jury, Charles Bouveyron, Frédéric Chazal, Aurélie Fischer et Clémentine Prieur qui m'ont fait l'honneur d'accepter d'évaluer ce travail. Un grand merci aux rapporteurs, Charles Bouveyron et Clémentine Prieur, d'avoir relu le manuscrit et de m'avoir fait part de leurs commentaires, ce qui me permet de proposer une version plus aboutie du manuscrit.

J'adresse aussi mes chaleureux remerciements à mes encadrants, Marc Glisse et Pascal Massart. Marc, j'ai pû bénéficier de tes larges connaissances, particulièrement en analyse topologique des données et en informatique, et admirer ta capacité à résoudre quasi-instantanément une grande variété de problèmes (même s'il a été frustrant de voir une difficulté qui m'occupait depuis plusieurs semaines s'effondrer par un contre-exemple que tu as construit en moins de 5 secondes). Pascal, j'ai bénéficié de tes connaissances encyclopédiques en statistique, ton flair mathématique et ton optimisme sans faille quant au fait que les données finiraient par donner raison aux approches développées dans cette thèse. Au delà de l'encadrement scientifique, j'ai bénéficié de votre soutien moral, de votre sympathie, de votre confiance et de vos conseils avisés pour faire mes premiers pas dans le monde de la recherche.

Ce travail ne serait rien sans la collaboration avec Metafora. Je remercie l'ensemble des salariés de l'entreprise avec qui j'ai eu la chance de travailler et de partager des moments agréables ces trois dernières années. Une part importante des contributions de cette thèse n'aurait jamais vu le jour sans cette relation. J'ai eu un accès facile à des données pour pouvoir tester et améliorer mes méthodes, et la possibilité de discuter avec des experts de la cytométrie. Je remercie en particulier Vincent Petit, dirigeant de la société, qui a été à l'origine de la collaboration et a tout mis en œuvre pour qu'elle porte ses fruits scientifiquement et industriellement. Je remercie aussi Baptiste Labarthe, responsable du pôle data science, pour sa disponibilité et son enthousiasme qui ont grandement facilité cette collaboration.

J'ai passé une part importante de ces quatres dernières années à l'Institut de Mathématique d'Orsay. C'est avec émotion que je vais quitter ce lieu, où j'ai eu la chance de côtoyer au quotidien des chercheurs aussi brillants que sympathiques, qui ont été mes professeurs puis mes collègues. Je dois beaucoup à eux et à l'ensemble du personnel de l'IMO qui œuvre au quotidien à faire de cet endroit un lieu de travail agréable.

J'ai eu la chance d'être membre de l'équipe Datashape. Je tiens ici à remercier tous les datashapeurs, d'aujourd'hui et d'hier, que j'ai eu le privilège de connaître. C'était pour moi un immense plaisir de faire parti d'une équipe réunissant des gens tous brillants, et dans laquelle règne un climat de décontraction et d'excellence sans pareil.

Un grand merci à tous les doctorants du LMO, de Datashape, et les anciens du M2 Stat-ML, avec qui j'ai partagé cette aventure et qui m'ont apporté un soutien moral capital. A ceux qui ont soutenu il y a déjà quelques temps, vous m'avez inspiré et démontré qu'il était possible de venir à bout de cette épreuve. A ceux qui franchissent la ligne d'arrivée avec moi, nous avons partagé nos doutes et nos réussites, cela m'a été d'une aide précieuse. A ceux qui soutiendront plus tard, je vous souhaite beaucoup de succès.

Je ne peux pas terminer ces remerciements sans citer ceux m'ont apporté une aide directe dans ce travail et que je n'ai pas encore nommé. Un grand merci à Etienne Lasalle d'avoir relu l'introduction. Un grand merci également à Jisu Kim pour ses explications concernant la convergence uniforme des estimateurs.

Enfin, un grand merci aux membres de ma famille, mes soutiens les plus fidèles depuis bientôt 27 ans.

Contents

Chapter 1

Introduction

Cette thèse est le fruit d'une collaboration entre l'équipe de recherche Datashape de l'Inria, spécialisée en analyse topologique de données, le laboratoire de mathématiques d'Orsay et plus particulièrement l'équipe probabilité et statistique et l'entreprise Metafora, qui développe des outils d'analyse du métabolisme cellulaire et des solutions de diagnostic innovantes à partir d'expériences de cytométrie en flux. La cytométrie permet d'effectuer des mesures sur un échantillon de cellules, par exemple à partir un prélèvement sanguin. Les mesures sont effectuées à l'échelle de la cellule individuelle et on mesure la présence de certaines protéines, appelées marqueurs membranaires. L'objectif est d'identifier des populations cellulaires. Dans un prélèvement sanguin on peut vouloir mesurer la présence de certaines cellules du système immunitaire. Parmi les applications, citons le suivi des patients séropositifs à l'aide du comptage des populations lymphocytaires [7]. La détection des populations est aujourd'hui encore largement réalisée manuellement.

Ce travail présente une approche de classification non supervisée automatique des données de cytométrie basée sur une modélisation originale. A travers la collaboration avec la société Metafora, nous avons cherché à comprendre les attentes des cytométristes. Nous avons conclu de ces échanges qu'il existait à la fois un besoin réel d'automatiser certaines tâches de traitement de données pour gagner du temps et éliminer certains biais des analyses manuelles. Ce besoin se fait d'autant plus ressentir qu'au cours des 30 dernières années, la dimension maximale des données produites en cytométrie a augmenté pour passer de 2 ou 3 au début des années 90 à plus de 40 aujourd'hui. Ce passage de données de faible dimension

à des données de dimension modérée induit une augmentation du temps d'analyse manuelle ainsi qu'une diminution de la robustesse des analyses. Des solutions automatiques existent, mais nous avons identifié certains freins à leur adoption. D'une part, la plupart des méthodes disponibles nécessitent de calibrer des hyperparamètres, tâche souvent difficile pour un public non expert de l'analyse de données. D'autre part, le caractère "boîte noire" de certaines approches ne permet pas d'envisager une utilisation pour des applications cliniques.

Il nous semblait alors pertinent de développer une solution algorithmique suffisamment simple pour pouvoir être intégrée dans une interface, avec un temps de calcul raisonnable pour des données de dimension modérée et intégrant une phase de réduction de dimension par l'apprentissage d'un modèle cohérent par rapport aux données. Notre choix s'est porté vers l'utilisation de l'algorithme de classification automatique ToMATo [14]. Il prend en entrée un jeu de données ainsi qu'une fonction de densité et détermine une classification non supervisée hiérarchique en temps quasi-linéaire. Il utilise des techniques d'analyse topologique des données et bénéficie des résultats théoriques de stabilité du domaine. Ce choix nous laisse une liberté sur la méthode d'estimation de densité. Nous avons dans un premier temps utilisé des estimateurs basés sur la distance à la mesure (DTM) [6], puis nous avons développé une approche pour l'estimation de densité sous l'hypothèse de structure d'indépendance, introduite par Lepski [42] et Rebelles [65]. Celle-ci consiste à supposer que les variables peuvent être partitionnées en paquets de variables indépendants. Ce modèle nous semble pertinent par rapport aux approches manuelles utilisées par les cytométristes, qui considèrent implicitement des structures de groupes dans les variables. Nous avons en particulier implémenté un algorithme d'estimation de densité qui apprend une structure d'indépendance à partir des données en dimension modérée et démontré une borne supérieur sur le risque de l'estimateur ainsi construit. Notre démarche a été guidée par la volonté de trouver un équilibre entre une méthodologie statistiquement pertinente et la possibilité d'application aux données réelles.

Cette introduction vise à situer le contexte scientifique et la contribution de cette thèse. En section 1.1 nous proposons un rapide historique de la biologie cellulaire dans lequel s'inscrit le développement de la cytométrie, que nous présentons plus en détail en section 1.2 en insistant sur le phénomène de montée en dimension. Ensuite, nous présentons le domaine de l'analyse des données (manuelle et automatique) de cytométrie en section 1.3 puis celui de l'estimation de densité sous des hypothèses d'indépendance en section 1.4. Enfin, nous détaillons les contributions de ce travail dans la section 1.5, en annonçant le plan du manuscrit.

1.1 Rapide historique des connaissances et des techniques en biologie cellulaire

La cytométrie est une technique permettant d'effectuer des mesures sur chacune des cellules contenues dans un échantillon (un prélèvement sanguin par exemple). Elle s'inscrit dans un cadre plus large d'évolutions techniques dont nous brossons ici succinctement le portrait.

La première observation d'une cellule vivante est attribuée au drapier néerlandais Antoni van Leeuwenhoeck en 1674. Elle résulte du perfectionnement des techniques de microscopie. Ses observations se limitent cependant à des cellules naturellement en suspension dans un corps liquide comme les spermatozoïdes ou les protozoaires dans l'eau douce. Au cours du 18^{e} et du 19^{e} siècle, le développement des techniques de coupe histologique, qui consiste à extraire des tranches suffisamment fines d'un organe pour pouvoir l'observer au microscope, permet de multiplier les observations de cellules.

Ces observations forment le substrat de la biologie cellulaire, théorie qui sera développée au cours du 19^{e} siècle et dont la paternité est accordée à Mathias Jakob Schleiden, Theodor Schwann et Walter Flemming. Retenons qu'à la fin du 19^{e} siècle, l'idée s'est imposée au sein de la communauté scientifique que la cellule est l'unité de base des organismes vivants et que son mode de prolifération est la reproduction via le mécanisme de division cellulaire.

Le 20^{e} siècle est le théâtre d'un accroissement spectaculaire dans la connaissance liée à la cellule. Cette évolution est liée au développement de la biologie moléculaire. Cette discipline a pour objectif de traduire les fonctions de la cellule au niveau de ses composants moléculaires. Parmi les apports de cette discipline, on peut citer la théorie fondamentale de la biologie moléculaire décrite par Francis Crick en 1958 [19]. Elle met en évidence le lien entre les différents types de macromolécules au sein de la cellule: l'ADN (acide désoxyribonucléique), l'ARN (acides ribonucléiques) et les protéines. L'ADN porte le patrimoine génétique de l'individu. Les protéines assurent les différentes fonctions de la cellule (comme la mobilité, le métabolisme ou la transmission de signaux cellulaires). Les ARN sont des copies de parties du code génétique, qui entrent en jeu dans la production de protéines et jouent un rôle d'intermédiaire entre ADN et protéines. Cette théorie est schématisée par la figure 1.1. L'autre fait marquant de ce siècle

est le développement d'outils techniques, notamment à partir de 1970, qui vont permettre un accroissement des mesures possibles et permettre d'envisager des applications cliniques. Un exemple est le séquençage de l'ADN, dont les premières descriptions furent données par Maxam-Gilbert [25] et Sanger [73] en 1977. Cette technique permet de déterminer la succession des bases A, C, G et T constituants le génome. C'est également pendant le 20^{e} siècle que naît la cytométrie, que nous décrirons plus tard dans cette introduction.

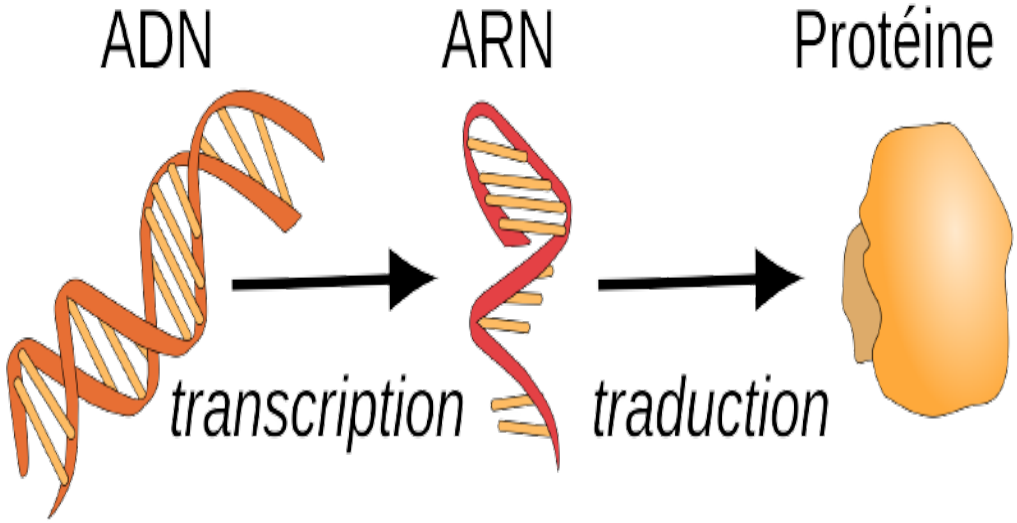


Figure 1.1: Illustration de la théorie fondamentale de la biologie moléculaire (Source : "Théorie fondamentale de la biologie moléculaire", Wikipedia, Wikimedia Foundation, 30 mars 2022)

Au 21^{e} siècle, les techniques de mesures continuent de se perfectionner, on peut s'en rendre compte en consultant l'article de revue [49] qui décrit l'évolution des techniques de séquençage d'ADN entre 2006 et 2016. Un autre fait marquant est la démocratisation des techniques liée à la baisse des coûts comme illustré dans le cas du séquençage de l'ADN par la figure 1.2. Ces évolutions s'accompagnent cependant de nouvelles problématiques, en particulier celle du traitement des données. La massification des données récoltées nécessite la mise en œuvre de méthodes informatiques adaptées. C'est dans ce contexte que se développe aujourd'hui la bio-informatique, champ de recherche multidisciplinaire à la croisée de l'informatique, de la statistique, des mathématiques, de la biologie et de la médecine. Les objectifs poursuivis par cette discipline très active sont,

entre autres, l'accroissement des connaissances induit par la possibilité d'obtenir de nouvelles mesures, l'accélération du processus global de recherche par la structuration et le partage de bases de données et le développement de nouveaux outils de diagnostic et de thérapie.

Dans cette thèse nous allons plus précisément nous intéresser à la cytométrie, une technique notamment utilisée pour l'identification de populations cellulaires par la détection de protéines particulières, les récepteurs membranaires.

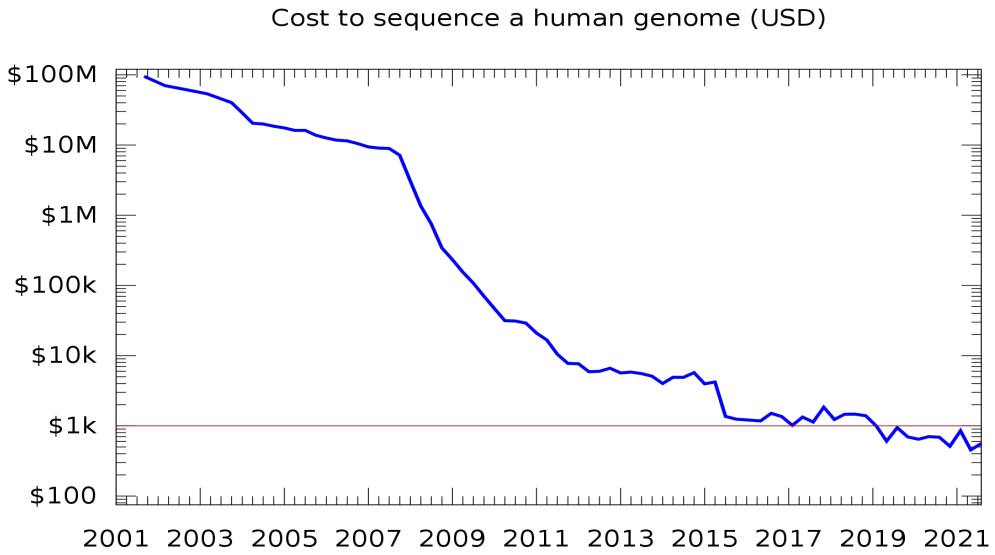


Figure 1.2: Démocratisation du séquençage ADN entre 2001 et 2021 (Source : "Séquençage de l'ADN", Wikipedia, Wikimedia Foundation, 27 août 2022)

1.2 Principe de la cytométrie

La cytométrie est une technique qui se développe au cours du 20^{e} siècle et poursuit son développement aujourd'hui, en phase avec les grandes évolutions que nous venons de présenter. Elle permet de mesurer la présence de protéines à la surface de cellules en suspension dans un fluide. On l'utilise en particulier pour traiter des échantillons sanguins. La cytométrie repose sur le couplage de trois sous-systèmes : un système fluidique qui permet de faire passer une à une les cellules

de l'échantillon dans un tube, un système analogique qui permet d'effectuer des mesures sur chaque cellule et un système numérique qui transmet les données acquises par le système analogique.

Les premiers cytomètres avaient pour fonction de compter les cellules d'une taille donnée. Le système analogique était alors un simple faisceau lumineux, obstrué au passage d'une cellule. La première description d'un tel système a été proposée en 1934 par Moldavan [53], mais il faut attendre 1947 pour voir la publication d'un article correspondant à la description du premier cytomètre effectivement réalisé [30]. Le schéma de fonctionnement de l'appareil est décrit par la figure 1.3.

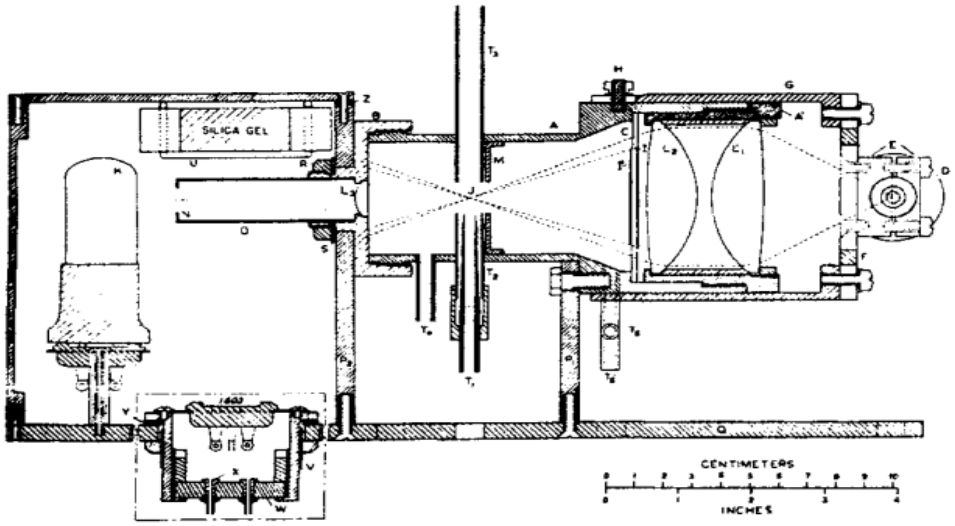


Figure 1.3: Schéma représentant le premier cytomètre opérationnel ([30])

Pendant la seconde moitié du 20^{e} siècle, la cytométrie en flux se développe. Il devient possible de mesurer la quantité de présence de certaines protéines, appelées récepteurs membranaires, à la surface des cellules. L'évolution majeure permettant ce développement est la mise au point de techniques de production d'anticorps monoclonaux, la première remonte à 1975 [38]. Un anticorps monoclonal est une molécule qui a la particularité de réagir au contact d'un antigène en se liant à celui-ci. Dans le cadre qui nous intéresse, ces antigènes sont les récepteurs membranaires. Le fonctionnement de la cytométrie en flux repose sur le principe de l'immunofluorescence. D'abord, une réaction chimique permet de

coupler chaque anticorps avec une molécule fluorescente appelée fluorochrome. Ensuite, si l'antigène visé est présent à la surface d'une cellule, le couple anticorps/fluorochrome vient se lier à celle-ci. Ce principe est illustré par la figure 1.4.

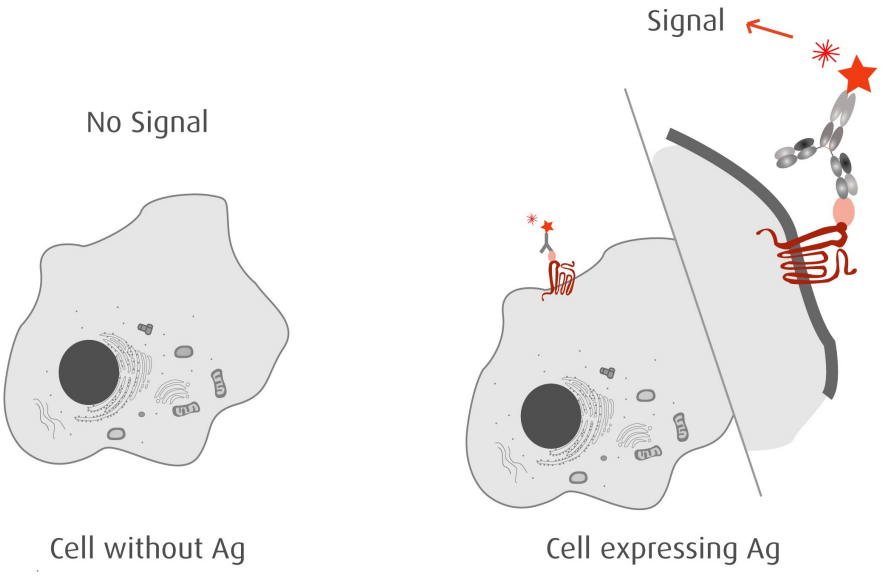


Figure 1.4: Principe de l'immunofluorescence

Le cytomètre en flux est un appareil dont le système analogique est un système optique permettant de caractériser le spectre d'émission lumineuse des cellules soumises à une irradiation par un laser. Les mesures correspondent à la quantité de flurochrome liée à la surface de la cellule et donc à la quantité de présence d'un antigène cible après préparation de l'échantillon selon le principe d'immunofluorescence. Le principe de fonctionnement d'un cytomètre en flux est illustré par la figure 1.5. Le système de mesure optique est constitué d'un ensemble de PMT (pour photomultiplier tube). Leur nombre correspond au nombre de variables mesurées par cellules. Les PMT nommées FSC (front scatter) et SSC (side scatter) mesurent des propriétés géométriques de la cellule (taille pour FSC et complexité pour SSC). Les autres PMT sont associés à des filtres passe-bande destinés à détecter la présence ou non de certains fluorochromes à la surface des cellules. L'augmentation du nombre de mesures par cellules est liée à l'augmentation du nombre de ces PMT. Le principal frein à leur multiplication est le phénomène de superposition des spectres d'émission des fluorochromes utilisés. Un fluorochrome destiné à émettre à une longueur d'onde à tendance à émettre aussi à des longueurs d'onde voisines et peut, selon le réglage, activer

d'autres PMT que celui avec lequel il est censé être couplé. Ce phénomène est inévitable et donne lieu à un prétraitement des données appelé compensation où l'expérimentateur corrige les signaux mesurés. Cependant, lorsque la finesse des bandes d'émission des fluorochromes et d'absorption des PMT est limitée on ne peut pas dépasser une certaine dimension.

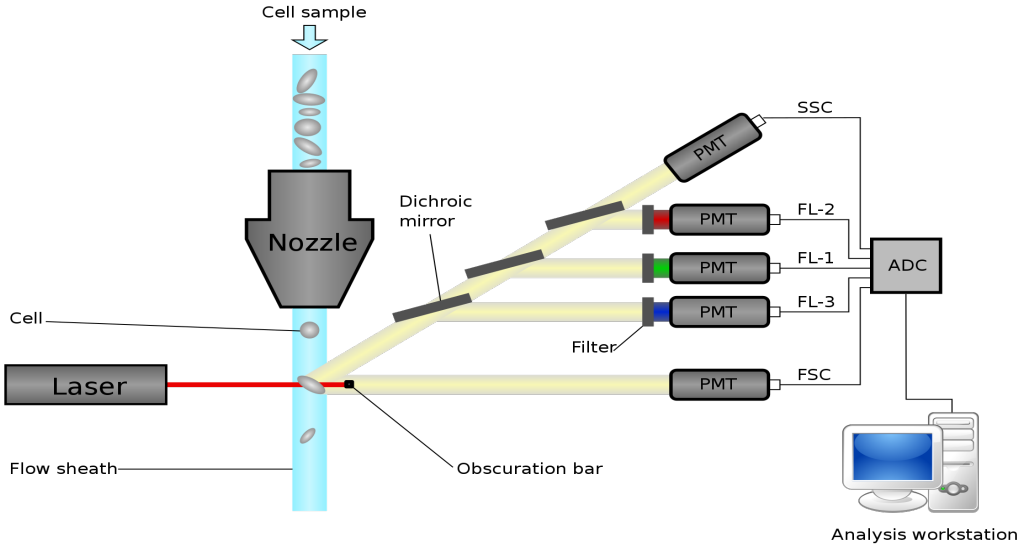


Figure 1.5: Principe de fonctionnement d'un cytomètre en flux (Source : "Flow Cytometry", Wikipedia, Wikimedia Foundation, 24 août 2022)

À partir des années 1990, le nombre de couleurs (c'est-à-dire de PMT qui ne sont ni FSC ni SSC) augmente petit à petit dans les expériences de cytométrie. On voit par exemple des expériences avec 5 couleurs en 1995 [69], 8 couleurs en 1997 [70], 11 couleurs en 2001 [20] et 17 couleurs en 2004 [61]. Autour de l'année 2010, deux révolutions viennent accélérer l'augmentation du nombre de couleurs. Ce sont les techniques de la cytométrie spectrale [67] et de la cytométrie de masse [3]. La cytométrie spectrale repose également sur le principe d'immunofluorescence, mais permet de recueillir l'intégralité du spectre d'émission des cellules plutôt que le signal en sortie des PMT. La cytométrie de masse se caractérise par le remplacement des fluorochromes par des isotopes de métaux non radioactifs, et les mesures sont liées à la masse atomique de ces métaux. Grâce à ces nouvelles techniques, on produit aujourd'hui des jeux de données de plus grande dimension (un exemple d'expérience avec 40 couleurs, donc 42 dimensions en cytométrie spectrale peut être trouvé dans [58] et avec 47 dimensions en cytométrie de masse

dans [68]).

Une fois les données récoltées, il reste à les traiter pour résoudre des questions biologiques. La section suivante précise les enjeux liés à la classification non supervisée des données de cytométrie.

1.3 Analyse des données de cytométrie

Dans la suite de cette introduction et le reste de la thèse, nous allons nous intéresser au problème de l'analyse des données. Nous ferons volontairement abstraction des considérations techniques sur la production des jeux de données et nous considérerons de manière indistincte les expériences de cytométrie en flux, spectrale ou de masse. De plus, dans nos applications, nous utiliserons des données publiques qui ont déjà été prétraitées par des experts du domaine. Nous adopterons la notation générique $X_1, \ldots, X_N$ pour désigner le jeu de données à disposition. N est le nombre d'événements mesurés par le cytomètre, c'est-à-dire le nombre de cellules. Pour i compris entre 1 et N, X_i est un vecteur dans $\mathbb{R}^d$, où d est le nombre de variables mesurées. Retenons que la valeur de N est typiquement de l'ordre de quelques dizaines ou quelques centaines de milliers et d peut atteindre des valeurs proches de 50. Les données qui nous intéressent sont de dimension modérée. Notre cadre d'étude diffère de la grande dimension, où d peut atteindre plusieurs centaines voir milliers et où on a parfois $N > d$. Nous ne sommes toutefois pas dans un régime où la dimension est assez faible pour s'affranchir de considération liées à la robustesse des méthodes face à la valeur de d.

Dans ce travail, nous nous concentrons sur la question de la classification non supervisée. L'objectif est d'identifier des populations cellulaires au sein de l'échantillon. Plus formellement, on cherche à définir une partition des indices $\{1, \ldots, N\}$. L'approche traditionnelle est celle du gating manuel, réalisée par la cytométriste. Plus récemment, des approches automatiques ont été proposées. Nous présentons ci-dessous ces deux approches.

Le gating manuel consiste à représenter les données par des nuages de points en deux dimensions en choisissant une paire de variables et à sélectionner une partie des données à l'aide d'un outil graphique. L'opération peut ensuite être répétée sur

les données extraites et, de manière récursive, on finit par identifier les populations cellulaires d'intérêt. L'enchaînement de ces étapes est appelé la stratégie de gating. La figure 1.6 représente une stratégie de gating pour l'identification de différents types de globules blancs. L'antigène CD45 est exprimé par les globules blancs, les cellules exprimant le flurochrome correspondant (AmCyan) sont réunies au sein d'une gate dans la vue de gauche. Ensuite, à l'aide de l'expression des antigènes CD14, CD15, CD3, CD19, CD4 et CD8 on identifie les sous-familles de globules blancs dans les trois autres vues. l'acronyme CD signifie cluster of differentiation, les CD sont des récepteurs membranaires caractéristiques de certaines cellules du système immunitaire. Il en a été découvert plus de 300 chez l'homme. Leur liste est tenue à jour par l'organisme HCDM (*Human Cell Differentiation Molecules*), on peut visiter le site internet https://www.hcdm.org/ pour plus d'informations.

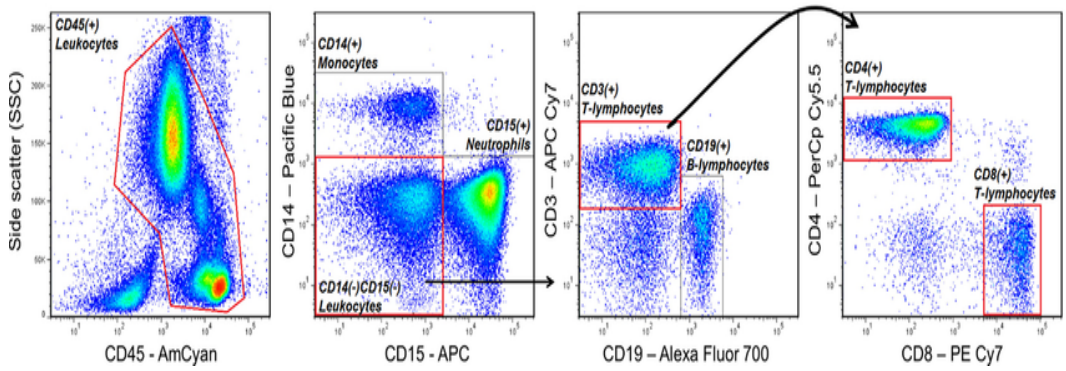


Figure 1.6: Illustration d'un gating manuel ([81])

La montée en dimension précédemment évoquée a un double impact sur la possibilité pour les cytométristes de proposer des stratégies de gating. D'un côté, en multipliant les mesures, on est en capacité de caractériser de plus en plus finement les populations cellulaires. D'un autre côté, en multipliant le nombre de stratégies de gating possibles, on perd de la robustesse et on augmente le temps nécessaire à l'analyse. Un exemple de stratégie de gating sur un fichier avec 40 variables est donné par la figure 1.7. Si l'on compare avec l'exemple décrit par la figure 1.6, on voit que le début de la stratégie est le même : on sélectionne les globules blancs comme les cellules exprimant CD45 (4ème vue de la première ligne, les 3 premières vues représentent des étapes de prétraitement destinées à éliminer les cellules mortes et les erreurs de mesures). Ensuite, on a un découpage très précis des différentes familles de globules blancs, représenté dans l'image par les rectangles de couleur qui représentent chacun l'identification d'une population cellulaire. On comprend facilement qu'une telle analyse est longue et qu'elle nécessite une forte

expertise biologique pour le choix des variables à chaque étape de la stratégie de gating.

Le choix des paires de variables dans la stratégie de gating n'est pas anodin. Certaines associations de variables sont plus pertinentes dans le carde de l'analyse. A l'inverse, certains marqueurs ne sont jamais utilisés conjointement Nous interprétons ce fait comme une modélisation implicite réalisée par le cytométriste. Par sa connaissance de la biologie sous-jacente, il identifie les populations d'intérêt sans avoir à considérer l'ensemble des paires de variables. Cependant, le choix des paires de variables n'est pas toujours évident et il est possible que deux experts adoptent deux stratégies différentes.

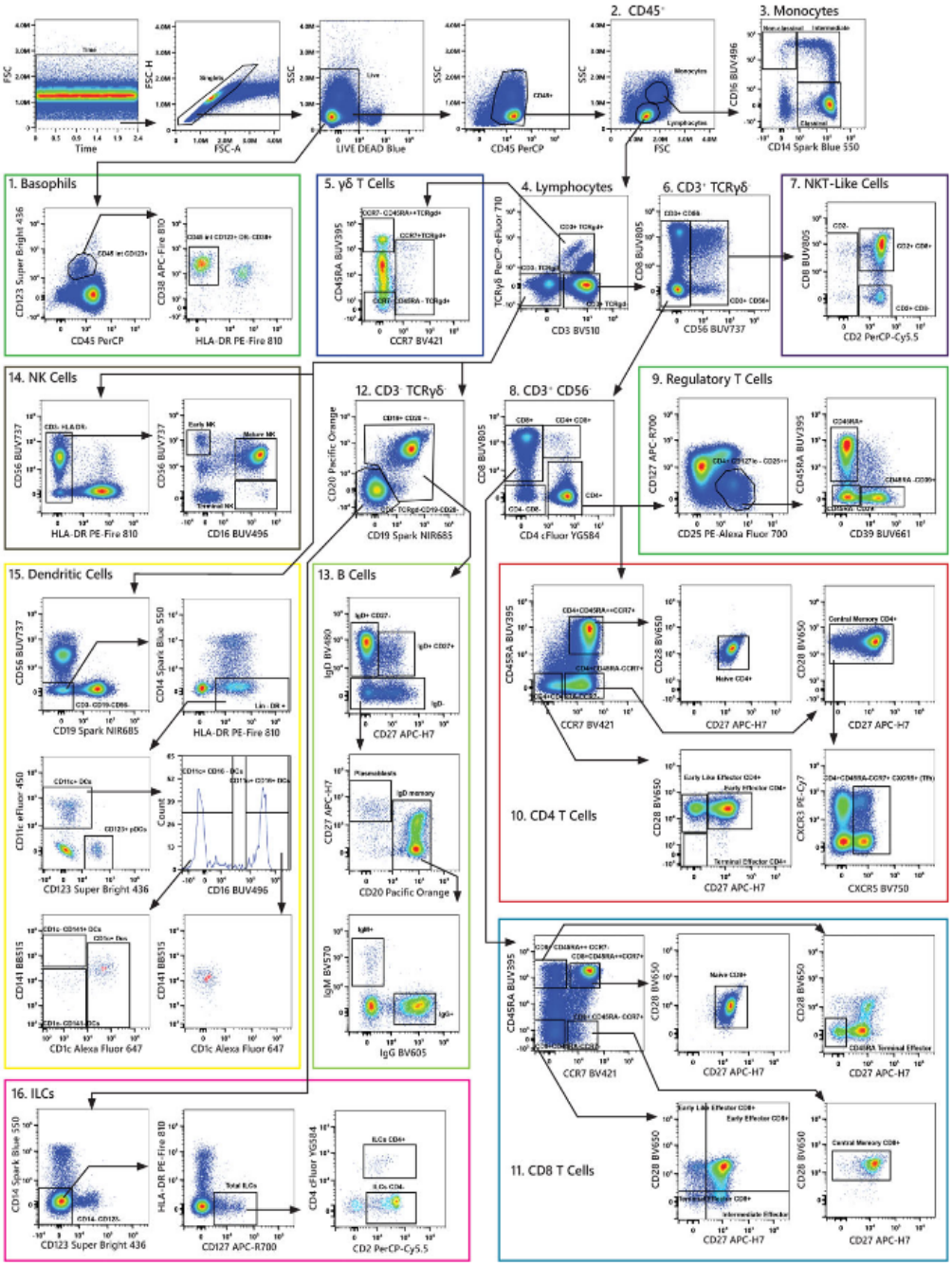


Figure 1.7: Gating manuel sur un jeu de données de dimension 40 ([58])

De nombreuses techniques de classification non supervisée existantes par ailleurs ont été adaptées aux données de cytométrie. Entre autres, on peut citer des approches par modèles de mélange (FLAME [63], flowCust [47]), par estimation de densité et détection de modes (Misty Mountains [76]), par l'algorithme des k-moyennes [48], par des cartes auto-organisatrices [39] (FlowKOH [80]) ou encore par partitionnement de graphe de voisinage (PhenoGraph [43]). Pour presque chacune de ces méthodes, il existe une variante ou la première étape consiste à exécuter un des algorithmes cités en réglant les paramètres pour obtenir un grand nombre de clusters, puis à exécuter un algorithme de concaténation des clusters ainsi obtenus. Cette approche permet en particulier de pouvoir obtenir des classifications hiérarchiques. Dans cette catégorie, on retrouve SWIFT [55] (basé sur un mélange de gaussiennes), X-shift [72] (basé sur une estimation de densité par plus proches voisins), flowMeans [1] (basé sur l'algorithme des k-moyennes), flowSOM [80] (basé sur des cartes auto-organisatrices).

Face à la prolifération de méthodes disponibles, il est devenu nécessaire de développer des moyens de comparaison. En 2011, une première étude comparée a été réalisée via le challenge FlowCAP-I (Flow Cytometry: Critical Assessment of Population Identification Methods) [74]. Cette étude a mis en avant le bon comportement des méthodes automatiques proposées pour des jeux de données contenant au plus 10 variables et au plus 100000 individus, sans toutefois montrer qu'une méthode particulière se montrait plus efficace que les autres. En 2016, Weber et Robinson [85] ont proposé un nouveau panel de données pour comparer des algorithmes de classification non supervisée, cette fois la dimension atteint 39 et le nombre d'événements quelques centaines de milliers. Sur ces données, [85] et [46] ont mis en avant les bonnes performances de l'algorithme FlowSOM [80]. Ces études utilisent des métriques basées sur le score F1 [34] permettant une quantification de l'adéquation entre un gating manuel et un algorithme de classification non supervisée.

Il est intéressant de comparer les résultats obtenus dans FlowCAP-I et Weber et Robinson. On note que parmi les algorithmes évalués dans les deux travaux, un certain nombre (SamSPECTRAL [88], FlowKOH [80] et FlowCust [47]), bien que performants lorsque la dimension est faible, perdent de leur efficacité comparativement à l'état de l'art lorsque la dimension augmente. Ce phénomène doit nous alerter. Il est possible que des algorithmes n'intégrant pas explicitement de phase de réduction de dimension demeurent performants pour les tailles de données actuellement produites en cytométrie. Cependant, rien ne garantie leur robustesse face à la montée en dimension. Certains travaux viennent aujourd'hui alerter sur

l'impact du fléau de la dimension en cytométrie [56], [75]. Ce phénomène, central en analyse de donnée et en statistique, explique la baisse de performance de la plupart des méthodes à mesure que la dimension augmente. C'est pour cela que nous avons pris la décision d'orienter une partie importante de ce travail de thèse sur le développement de techniques s'intégrant dans une procédure de classification non supervisée et faisant apparaître de manière explicite une réduction de la dimension.

Au-delà de la quantification de la performance des algorithmes de classification non supervisée par le score F1, il faut prendre en compte les attentes des cytométristes pour définir des approches pertinentes. On constate en effet aujourd'hui un écart important entre l'état de l'art de la recherche en analyse automatique de données de cytométrie et la pratique quotidienne des cytométristes, qui continuent majoritairement à utiliser des méthodes de gating manuel pour classifier leurs données. Nous identifions deux axes de réflexions pour comprendre ce phénomène.La première difficulté vient de la possibilité de rendre les algorithmes utilisables simplement par les cytométristes. La plateforme Cytobank [17] permet aux utilisateurs d'appliquer simplement des algorithmes d'analyse de données sur les données de cytométrie. Pour la classification non supervisée, c'est l'algorithme flowSOM qui a été retenu [23]. Cependant, flowSOM, comme la plupart des algorithmes précédemment cités, nécessite la calibration d'hyperparamètres, en particulier la dimension de la carte auto-organisatrice. Nous pensons que c'est ce qui freine son adoption par la communauté des cytométristes, qui n'ont pas forcément l'expertise requise pour utiliser ces méthodes. L'autre difficulté vient du fait qu'à long terme, ce sont les applications cliniques qui sont visées. On peut anticiper qu'un algorithme destiné à fournir une aide au diagnostic doive offrir certaines garanties de robustesse et de reproductibilité. La sortie de FlowSOM dépend d'une phase d'initialisation aléatoire, nous pensons que cet aspect peut être un frein pour l'application de l'approche dans un contexte industriel.

Dans notre démarche, nous avons choisi de nous concentrer sur une approche du partitionnement basé sur une estimation de densité non paramétrique. Dans la section 11.13 son livre de référence publié en 1975 [32], Hartigan définit les classes d'un partitionnement comme "des régions de l'espace de forte densité séparées les une des autres par des régions de faible densité". Ce point de vue nous semble intéressant dans un contexte multidisciplinaire, la notion de densité de probabilité étant assez largement répandue parmi les communautés scientifiques contrairement à des connaissances plus spécifiques aux statistiques ou à l'analyse de donnée qui sont à la base de certains algorithmes. De plus, à condition de calibrer a

priori les paramètres de l'estimation de densité, cette approche évite une étape d'initialisation aléatoire.

1.4 Estimation de densité

Notre approche repose sur l'enchaînement de deux étapes : une étape d'estimation de densité sur les données d'entrée et une étape de classification non supervisée hiérarchique déterministe, basé sur l'utilisation de l'algorithme ToMATo [14]. Plus précisément, ToMATo prend en entrée le jeu de données $X_1, \ldots, X_N$ et une fonction densité $\hat{f}$ définie en chacun des points : $\hat{f}(X_1), \ldots, \hat{f}(X_N)$. La manière dont $\hat{f}$ est calculée n'importe pas dans cette procédure, ce qui nous laisse une liberté sur la méthode d'estimation de densité. Nous avons trouvé pertinent de chercher à implémenter un estimateur de densité prenant en compte un modèle d'indépendance entre variables, inspirés par ce qui est fait en analyse manuelle des données de cytométrie où un modèle implicite permet de concevoir une stratégie de gating. Dans notre cas, plutôt que de s'appuyer sur une connaissance métier, nous cherchons à automatiser l'apprentissage du modèle à partir des données. Dans cette section, nous introduisons les éléments d'estimation de densité qui seront utiles à notre démarche.

L'estimation de densité est un sujet classique en statistique. Nous adoptons le point de vue fréquentiste : on suppose que les données $X_1, \ldots, X_N$ ont été tirées aléatoirement et de manière indépendante selon une loi de probabilité admettant une densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$. Nous nous plaçons aussi dans le cadre non paramétrique : la fonction f n'est pas caractérisée par un nombre fini de paramètres réels, mais appartient à un espace fonctionnel plus vaste. L'estimation de densité consiste à proposer un estimateur $\hat{f}$ à partir des données. Une manière de quantifier la difficulté du problème d'estimation de densité est de considérer le risque minimax. On considère une classe de fonction $\mathcal{F}$, appelée modèle et on suppose que f appartient à $\mathcal{F}$. Soit ℓ une fonction de perte sur l'espace des fonctions de densité. Étant donné un estimateur $\hat{f}$, on définit son risque maximum sur $\mathcal{F}$ par $\mathcal{R}(\hat{f}) = \sup_{f \in \mathcal{F}} \mathbb{E}\left[\ell(f, \hat{f})\right]$, où l'espérance est prise selon la loi des réalisations $X_1, \ldots, X_N$. Enfin, le risque minimax est défini comme le risque de l'estimateur avec le plus faible risque maximal. Les pertes classiques sont les pertes L_p avec $1 \leq p < \infty$: $L_p(f, g) = \left(\int (f-g)^p\right)^{1/p}$, la perte L_∞ : $L_\infty(f, g) = \|f - g\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x) - g(x)|$ ou la perte de Kullback-Leibler

(KL): $\mathrm{KL}\,(f\|g) = \int \log(f/g)f$.

En général, on ne cherche pas à calculer explicitement le risque minimax mais plutôt à donner son comportement asymptotique lorsque N tend vers l'infini. En estimation de densité et pour des pertes classiques, on obtient souvent des vitesses de convergences en $(1/N)^\alpha$, où α est lié à la régularité de la classe $\mathcal{F}$. Conceptuellement, plus α est élevé, plus le problème d'estimation peut être considéré comme simple (du moins asymptotiquement). Une introduction au sujet peut être trouvée dans [77]. Pour une revue complète et à jour des résultats en perte L_p, on peut se référer à [28] pour l'obtention des bornes inférieures, qui caractérisent la complexité par des critères de distinguabilité entre hypothèses du modèle et à [29] pour l'obtention des bornes supérieures qui nécessitent la construction d'estimateurs optimaux.

À titre d'exemple, si $\mathcal{F}$ est une classe de densités sur $\mathbb{R}^d$ de régularité de Hölder β (la définition précise n'est pas nécessaire ici, retenons simplement que plus β est élevé, plus les fonctions de la classe auront des dérivées d'ordre important bornées, et seront donc plus régulières), alors le taux de convergence du risque minimax pour la perte L_2 au carré est $(1/N)^{\frac{2\beta}{2\beta+d}}$ [33]. On remarque que plus le modèle se limite à des fonctions régulières (plus β est élevé), plus l'estimation est facile. À l'inverse, plus la dimension d augmente, plus le problème devient difficile, ce phénomène est une manifestation du fléau de la dimension dans ce contexte particulier.

La régularité des densités à estimer est un paramètre en général inconnu, on a alors recours à des estimateurs adaptatifs. Ceux-ci sont construits pour être optimaux sur une famille de modèles. Une technique qui a du succès pour les applications est celle de la validation croisée. Le principe repose sur la séparation de données en deux ensembles disjoints, l'un d'entraînement et l'autre de test. L'ensemble d'entraînement est utilisé pour construire les estimateurs correspondants aux différentes valeurs possibles des paramètres à calibrer et l'ensemble de test sert à mesurer la qualité de ces estimateurs par un critère empirique. On parle de validation croisée lorsque cette opération est répétée avec plusieurs séparations entraînement/test et le paramètre sélectionné est celui qui performe le mieux en moyenne. L'approche la plus exhaustive est le leave-p-out, on choisit une taille d'échantillon test et on fait toutes les séparations correspondantes, il y en a p parmi N. Moins vorace en ressources de calcul et plus utilisée, l'approche v-fold consiste, pour un entier v (typiquement 5 ou 10), à séparer les données en v échantillons de taille homogène. Ensuite, on utilise alternativement l'un des

échantillons comme test et les $v-1$ restantes pour l'entraînement. Une revue des méthodes de validation croisée peut être trouvée dans [2].

La dimension des données est connue, mais handicapante car elle ralentit le taux de convergence. La prise en compte d'un modèle traduisant des relations d'indépendances entre variables peut permettre de réduire l'impact négatif de la dimension. Un tel modèle peut être définit à partir de la notion de modèle graphique non orienté (une introduction au domaine se trouve dans [27] et un traitement plus en profondeur dans [82]). Soit un graphe $G = (V, E)$, où les sommets V représentent les variables $\{1, \ldots, d\}$ et E un ensemble d'arêtes (c'est-à-dire de paires de sommets). G est un modèle graphique pour la variable aléatoire sur $X = (X^1, \ldots, X^d)$ si la condition suivante d'indépendance conditionnelle est vérifiée :

$$(i, j) \notin E \Rightarrow X^i \perp\!\!\!\perp X^j | (X^k)_{k \notin (i,j)}. \tag{1.4.1}$$

Elle se lit de la manière suivante : si les sommets i et j ne sont pas reliés, alors les variables X^i et X^j sont indépendantes conditionnellement à toutes les autres variables. Les contraintes sur le modèle graphiques imposent une structure particulière à la densité de X. Une telle structure peut permettre de s'affranchir du fléau de la dimension. Cependant, apprendre un modèle graphique est difficile en général. Un résultat est que si G est un modèle graphique pour X, en notant $\mathcal{C}$ l'ensemble des cliques de G (c'est-à-dire l'ensemble des familles de sommets complètement connectées), il existe une collection de fonctions positives $(\psi_C)_{C \in \mathcal{C}}$ telles que la densité f de X s'écrit sous la forme :

$$f(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C) \tag{1.4.2}$$

pour tout x dans $\mathbb{R}^d$, où x_C est le la projection de x sur l'espace vectoriel engendré par les vecteurs unitaires $(e_i)_{i \in C}$. Z est une constante de renormalisation. Comme remarqué dans la section 2.1.2 de [82], les fonctions $(\psi_C)_{C \in \mathcal{C}}$ n'ont pas de lien clair avec les densités marginales de f. Estimer le modèle graphique G à partir d'un échantillon est une tâche trop difficile si on ne restreint pas la classe de graphes admissibles pour G. À notre connaissance, deux modèles ont été étudiés dans un contexte non paramétrique : le cas où G est une forêt et celui de la structure d'indépendance. Nous les présentons ci-dessous.

Une forêt est un graphe ne contenant pas de cycle. Si on considère que X admet une forêt G pour modèle graphique, alors la densité f peut s'écrire à partir

des densités marginales unidimensionnelles et bidimensionnelles de f :

$$f(x) = \prod_{(i,j)\in E} \frac{f_{\{i,j\}}(x_i, x_j)}{f_{\{i\}}(x_i) f_{\{j\}}(x_j)} \prod_{k=1}^{d} f_{\{k\}}(x_k). \tag{1.4.3}$$

Ce modèle a été étudié dans [45]. Les auteurs de ce travail présentent un algorithme permettant d'estimer une forêt et un estimateur associé à partir des données, appelé FDE (Forest Density Estimation). Il consiste à estimer les marginales de f de dimension 1 et 2, puis de déterminer l'estimateur de la forme 1.4.3 correspondant le mieux aux données selon un critère empirique. Le théorème 9 de [45] montre que si X admet une forêt pour modèle graphique, alors la vitesse de convergence de la procédure en perte de Kullback-Leibler (KL) est liée à la vitesse d'estimation de densité en dimension 2 et non en dimension d. Ce résultat met en lumière que l'hypothèse structurelle sur G constitue un remède au fléau de la dimension.

La structure d'indépendance, introduite par Lepski [42] dans le contexte de l'estimation en norme L_∞ et étudiée ensuite par Rebelles [65] pour les pertes L_p avec $1 \leq p < \infty$ consiste à considérer que les variables peuvent être décomposées en blocs indépendants. Soit k un entier inférieur à d. On note Set_d^k l'ensemble de sous-ensembles de $\{1, \ldots, d\}$ de taille au plus k, nous parlerons de paquets de variables, Part_d^k l'ensemble des partitions de $\{1, \ldots, d\}$ que l'on peut former à partir de paquets contenus dans Set_d^k. Le cardinal de Set_d^k sera noté S_d^k et celui de Part_d^k sera noté B_d^k. Le modèle de structure d'indépendance correspond à l'ensemble des lois de probabilité sur $\mathbb{R}^d$ dont la densité peut s'écrire comme un produit de densités marginales. Une définition formelle du modèle de structure d'indépendance d'ordre k est donné par

$$\mathcal{D}_d^k = \left\{ f : \mathbb{R}^d \mapsto \mathbb{R} \mid \exists \mathcal{P} \in \mathrm{Part}_d^k : f(x) = \prod_{S\in\mathcal{P}} f_S(x_S) \right\}. \tag{1.4.4}$$

Dans une perspective de modèle graphique, cela correspond à considérer que X admet un graphe G pour modèle graphique, où G est composée de composantes disjointes de taille au plus k. Il a été montré dans [42] et [65] que le taux de convergence du risque minimax sous une telle hypothèse était lié à k plutôt qu'à la dimension ambiante d. En particulier, dans [65] il est montré que pour la perte L_2 au carré, le taux de convergence du modèle contenant les densités satisfaisants une hypothèse de régularité Hölder β et une structure d'indépendance d'ordre k est $(1/N)^{\frac{2\beta}{2\beta+k}}$. L'effet du fléau de la dimension est là aussi atténué et le taux de convergence ne dépend plus que de la taille du plus gros bloc de variables dans la

décomposition en structure d'indépendance. Des estimateurs minimax adaptatifs en la régularité et en la structure d'indépendance ont été proposés par Lepski et Rebelles. Cependant, leur définition fait apparaître la résolution de problèmes d'optimisations sur Part_d^k par méthode exhaustive. Or la taille de Part_d^k devient rapidement trop grande lorsque d grandit pour pouvoir résoudre numériquement ce problème, même si l'on se limite à $k = 2$.

Nous avons choisi d'explorer les possibilités d'utilisation du modèle de structure d'indépendance. Notre intuition est que ce modèle est bien adapté à la cytométrie. En effet dans l'analyse par gating manuel, les variables d'entrée sont utilisées par bloc afin de classifier les données. Il ne semble pas absurde de penser que certains groupes de variables vont coder une information sur les cellules indépendantes d'autres groupes de variables. De plus, si une telle décomposition en structure d'indépendance permet d'apporter une information pertinente, elle constitue une information qualitative d'intérêt et simple à interpréter pour le cytométriste.

1.5 Contributions

La première contribution de cette thèse est de proposer un algorithme de classification non supervisée pour les données de cytométrie. L'algorithme développé est appelé CyToMATo (Cytometry Topological Mode Analysis Tool), il est présenté dans le chapitre 2. Son fonctionnement est basé sur l'utilisation de l'algorithme ToMATo (Topological Mode Analysis Tool) [14], qui utilise des techniques issues de l'analyse topologique des données pour proposer une classification non supervisée hiérarchique des données à partir d'une estimation de densité. Nous utilisons un estimateur de densité basé sur la notion de distance à la mesure [6] et montrons en analysant les performances en termes de score F1 sur les données du challenge FlowCAP-I que les résultats sont peu sensibles au choix précis des hyperparamètres à condition d'utiliser le logarithme de la densité au lieu de la densité elle même. Cette analyse nous permet de définir l'algorithme CyToMATo (algorithme 1), qui présente l'avantage de ne nécessiter aucun calibrage d'hyperparamètres et peut donc facilement être intégré dans une interface utilisable par des cytométristes qui ne sont pas experts des algorithmes d'analyse de données. L'algorithme est ensuite validé sur les données de cytométrie de masse étudiées par Weber et Robinson. Il ressort que les performances de CyToMATo sont légèrement inférieures

mais comparables à l'état de l'art, représenté ici par l'algorithme FlowSOM, tout en ayant l'avantage de ne pas nécessiter la calibration d'hyperparamètres. À ce stade de notre travail, nous avons proposé une solution algorithmique satisfaisant nos attentes en termes de performance et de capacité d'intégration dans un logiciel simple d'utilisation. Cependant, nous avons laissé de côté la question de la montée en dimension. La réduction de dimension pour les données de cytométrie est aujourd'hui souvent réalisée avec UMAP (Uniform Manifold Approximation and Projection) [51]. Cet algorithme permet de projeter non linéairement les données initiales dans un espace de dimension inférieure (en pratique 2 ou 3 pour pouvoir être visualisable). Nous montrons aussi dans ce chapitre que l'application de CyToMATo aux données réduites obtenues avec UMAP permet d'obtenir une légère amélioration des performances de clustering. Cette observation nous conforte dans l'idée que l'intégration d'une étape de réduction de dimension peut être pertinente pour le problème de la classification non supervisée des données de cytométrie. Cependant, UMAP a le désavantage d'agir comme un algorithme de type boîte noire. À la réduction de dimension effectuée n'est pas associée un descripteur simple qui permettrait d'expliquer le résultat obtenu. Nous avons alors décidé de réfléchir à la question de l'intégration de la réduction de dimension dans la phase d'estimation de densité de CyToMATo par l'apprentissage d'un modèle de structure d'indépendance.

La seconde contribution, présentée au chapitre 3 consiste en l'écriture d'un algorithme d'estimation de densité prenant en compte une éventuelle structure d'indépendance, ISDE (Independence Structure Density Estimation) (algorithme 2). Il prend en entrée un paramètre k et une méthode d'estimation de densité multidimensionelle pour l'estimation des densités marginales et retourne une partition $\hat{\mathcal{P}}$ composé de paquets de variables de taille au plus k et un estimateur $\hat{f}_{\hat{\mathcal{P}}}$ de la forme $\hat{f}_{\hat{\mathcal{P}}} = \prod_{S\in\hat{\mathcal{P}}} \hat{f}_S$ où $\hat{\mathcal{P}}$ et $(\hat{f}_S)_{S\in\hat{\mathcal{P}}}$ sont estimés à à partir des données. Les estimateurs proposés par Lepski et Rebelles dans leurs travaux sur la structure d'indépendance ne peuvent pas être directement implémentés car ils nécessitent de résoudre des problèmes d'optimisation exhaustifs sur Part_d^k. Notre approche repose sur l'écriture du problème d'estimation de densité en utilisant la perte de Kullback-Leibler et son pendant empirique, la log-vraisemblance. On fait ainsi apparaître des logarithmes de densités produits, et donc des sommes de logarithmes de densités marginales. Cette réécriture du problème d'optimisation comme un problème additif permet de procéder en deux temps. On commence par estimer séparément les densités marginales $(\hat{f}_S)_{S\in\mathrm{Set}_d^k}$ puis on trouve la manière optimale de les combiner en maximisant la somme de leur log-vraisemblance sous contrainte d'obtenir une partition des variables. En pratique les N données sont séparées en deux : m données servent à l'estimation des densités marginales et n au choix

de la partition. Dans le cadre de l'estimation de densité par noyaux gaussiens, nous tirons parti de l'implémentation proposée par la librairie KeOps [12] qui permet une accélération des calculs sur carte graphique. Pour trouver la meilleure partition comme combinaison d'estimateurs marginaux définis sur les paquets de variables, nous utilisons une approche par programmation linéaire et notamment le package PulP [52]. L'utilisation de ces outils nous fait gagner un temps de calcul important et nous permet d'envisager d'utiliser ISDE sur des données de cytométrie. Nous montrons aussi la pertinence de ISDE sur données simulées vérifiant la structure d'indépendance. ISDE est dans ce cas capable de détecter la structure d'indépendance et offre de meilleures performances que d'autres estimateurs de densité n'intégrant pas cette hypothèse (en particulier un simple estimateur à noyau en dimension ambiante) en terme de log-vraisemblance empirique sur données test.

Se pose alors la question de la pertinence statistique de l'estimateur ainsi construit. Pour y répondre, nous établissons dans le chapitre 4 une borne supérieure en grande probabilité sur le risque $\mathrm{KL}\left(f \| \hat{f}_{\hat{\mathcal{P}}}\right)$, où f est la vraie densité des données et $\hat{f}_{\hat{\mathcal{P}}}$ la densité obtenue par ISDE avec l'utilisation d'un estimateur à noyau gaussien pour l'estimation des marginales. Le résultat principal de ce chapitre est le théorème 4.4.1

Théorème 4.4.1

Si f satisfait une structure d'indépendance $\mathcal{P}_*$ d'ordre k^*, est à support dans $[0,1]^d$, que ses marginales sont de régularité Hölder $\beta \in]0,2]$, et qu'il existe $A > 0$ tel que le logarithme des marginales $(f_S)_{S\in\hat{\mathcal{P}}}$ est majoré par $A|S|$, alors si $k \geq k^*$, pour un bon choix d'estimateurs marginaux, il existe deux constantes C_1 et C_2 telles que, avec probabilité (1 - 2/m)(1-2/n)

$$\begin{aligned}\mathrm{KL}\left(f \| \hat{f}_{\hat{\mathcal{P}}}\right) \leq & C_1 \sqrt{\log m + \log\left(S_d^k\right)} \left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+k}} \\ & + C_2 \sqrt{\log n + \log\left(S_d^k\right)} \frac{k}{\sqrt{n}}. \end{aligned} \tag{1.5.1}$$

Nous tirons deux conclusions à partir de cette borne. Premièrement, l'exposant $\beta/(2\beta+k)$ montre que limiter la taille maximale des paquets de variables permet de limiter les effets du fléau de la dimension. Deuxièmement, la présence de la quantité $\log(S_d^k)$ dans les constantes traduit la réduction de complexité du problème

d'optimisation liée à l'écriture du problème d'optimisation de la log-vraisemblance comme un problème additif. Une approche exhaustive ferait apparaître la quantité $\log(B_d^k)$. Notre borne est adaptative en la structure d'indépendance à condition de connaître un majorant de la taille du plus grand paquet dans la décomposition de f en structure d'indépendance. Elle n'est cependant pas adaptative en la régularité qui est supposée connue dans l'établissement de la borne supérieure. L'estimateur utilisé diffère de celui utilisé en pratique, il sert ici à montrer la pertinence théorique de l'approche d'ISDE en autorisant l'utilisation de certains paramètres du modèle qui sont inconnus. En pratique, nous utilisons des techniques de validation croisée pour nous adapter à ces paramètres. Ces restrictions et les hypothèses fortes sur la vraie densité viennent de difficultés techniques liées à l'étude du problème d'estimation de densité en perte de Kullback-Leibler, plus délicate qu'en perte L_p. Cela est dû à la présence du logarithme dans la définition qui peut poser un problème si le ratio $f/\hat{f}$ n'est pas borné. La solution retenue est de s'assurer s'assurer que $\hat{f}$ ne s'approche pas arbitrairement près de zéro en supposant que les marginales f_S sont inférieurement bornées et en construisant les estimateurs marginaux $\hat{f}_S$ tels que les quantités $\|f - \hat{f}\|_\infty$ soient contrôlées.

Enfin, dans le chapitre 5, nous montrons l'intérêt de l'algorithme ISDE sur des données de cytométrie, celles de Weber et Robinson en particulier. Premièrement, nous montrons que ISDE offre de meilleurs résultats en termes de log-vraisemblance sur données test que d'autres estimateurs de densité adaptés à ce cadre. Nous étudions ensuite sur ces mêmes jeux de données l'espace des estimateurs reposant sur une structure d'indépendance en fonction de la partition considérée. Nous mettons en avant que la topologie induite par la distance edit ([9]) sur l'espace des partitions semble pertinente dans le sens où plus une partition est éloignée pour cette distance de la partition obtenue par ISDE, plus la performance en termes de log-vraisemblance sur données test se détériore. Cela nous conforte dans l'idée que la partition particulière de variables obtenue avec ISDE a un sens et peut être interprétée d'un point de vue biologique. Enfin, nous mettons en lumière que modifier CyToMAto en intégrant la partition des variables obtenue par ISDE permet d'améliorer légèrement les scores F1 moyens obtenus avec CyToMATo au chapitre 2. Cependant, il ne faut pas surinterpréter ce gain en terme de score F1, les résultats sous ce critère restent comparables. Malgré tout, il est intéressant de constater que les performances ne se détériorent pas, ce qui aurait pu se produire, car la structure d'indépendance est une hypothèse forte et a priori peut conduire à une perte d'information et donc une baisse des performances. Nous avons aussi extrait des données une information qualitative précieuse pour le biologiste via la partition des variables.

Deux annexes se trouvent à la fin de ce manuscrit. L'annexe A complète le chapitre 3 et montre les performances d'ISDE quand il est utilisé sur des données gaussiennes. Dans ce cadre, un algorithme a déjà été mis au point par Devijver et Gallopin [22], et nous montrons qu'ISDE permet d'obtenir des performances comparables. L'annexe B complète le chapitre 4 et présente une étude quantifiée du biais du modèle d'indépendance lorsque les données sont distribuées selon une loi normale multidimensionnelle. Nous quantifions en particulier dans ce cas l'erreur commise en considérant un modèle de structure d'indépendance quand la loi d'entrée présente des groupes de variables faiblement corrélés et lorsque le paramètre choisi de taille maximale des groupes est inférieur à la taille réelle des blocs de variables. Ces annexes apportent des compléments à certains aspects présentés au cours de cette thèse mais ne sont pas directement pertinentes dans le cadre de la modélisation des données de cytométrie, ne serait-ce que parce que les variables gaussiennes multidimensionnelles sont unimodales alors que la présence de différentes populations cellulaires dans un échantillon induit une distribution multimodale des données. Nous pensons cependant que le lecteur intéressé y trouvera des résultats complémentaires qui pourraient s'avérer utiles pour d'autres applications.

Chapter 2

CyToMATo: a hierarchical clustering algorithm tailored for cytometry data

Context In recent years, clustering algorithms in the context of cytometry data has been a topic of great interest. Various approaches were proposed, mostly adaptation of methodology available in the wider clustering literature to the specificity of cytometry data. In 2011, a first effort to benchmark the teeming approaches to automated classification of cytometry data was made. The FlowCAP-I challenge consisted in comparing algorithms on a defined set of datasets through the F1 score. Since this seminal work, other benchmark analysis was designed, incorporating new algorithms and type of datasets, particularly with higher dimensions. These works led to a consensus that some approaches are generally more efficient than others and furnish a baseline to assess the potential of a new clustering method, in terms of methodology, data and evaluation criteria. However, from our point of view, the existing clustering algorithms in the domain suffer from two main drawbacks: they rely on hardly tuneable parameters for a non-expert and do not address the curse of dimensionality.

Our contribution In this context, we propose a new approach, CyToMATo. It is a hierarchical version of the clustering algorithm ToMATo (Topological Mode Analysis Tool) [14] combined with a log-density estimation step using Distance-

to-Measure (DTM) approach [6]. We show that CyToMATo is not much sensitive to the choice of the hyper-parameters, and then it can be defined as a parameter-free algorithm, which is of great interest in an industrial context. This work has led to the writing of a patent. Here we prove that it compares with state-of-the-art clustering algorithms for cytometry data, following the methodology of [85] to evaluate clustering. We also show that a dimensionality reduction step with UMAP as a preprocessing of the data leads to an improvement in the performance. This encourages us to incorporate dimensionality reduction scheme in our procedure. However, the functioning of UMAP is quite opaque. For this reason, in the next chapters, we will define an original way to incorporate dimensionality reduction in the density estimation step via the Independence Structure (IS) model.

Organization of the chapter The chapter is organized as follows. In section 2.1 we review the functioning of ToMATo and explain how it can be turned into a hierarchical clustering algorithm. In section 2.2 we introduce the DTM-based density estimation. In section 2.3 we introduce CyToMATo, a hierarchical clustering algorithm based on ToMATo and a DTM-density estimation and show that a fixed value for hyper-parameters lead to performance not far from state-of-the-art for high dimensional cytometry data. Then in section 2.4 we show the potential of improvement induced by adding a dimensionality reduction step with the algorithm UMAP (Uniform Manifold Approximation and Projection).

2.1 ToMATo

We consider the context of having N data $X_1, \ldots, X_N$ in $\mathbb{R}^d$. Our goal is to partition the indices $\{1, \ldots, N\}$ into a collection of clusters $C_1, \ldots, C_K$. ToMATo (Topological Mode Analysis Tool) is a clustering algorithm introduced in [14], it aims to cluster data based on the topology of the upper-level sets of the density function that must be first estimated at the data points. In this section, we ignore the density estimation step and consider that data comes with corresponding density, we will see later how to estimate the density.

2.1.1 Description of the algorithm

Modes and upper-level sets' topology Let g be a function defined over the sample $X_1, \ldots, X_N$. In this presentation, we will consider a one dimensional example, illustrated by figure 2.1.

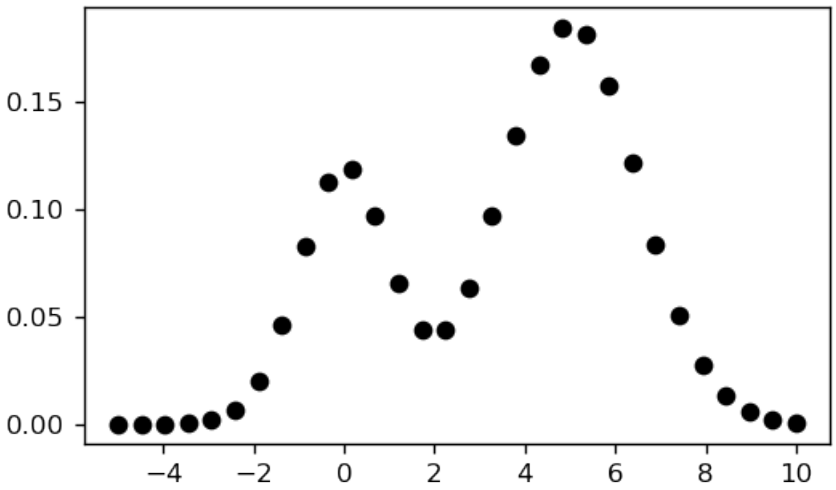


Figure 2.1: Illustrative dataset, the values of $X_1, \ldots, X_N$ are indicated on the x-axis and the values of $g(X_1), \ldots, g(X_N)$ on the y-axis

We aim to study the modes and the topology of the upper-level sets of g, thought as a continuous function only known over the point cloud $X_1, \dots, X_N$. Our intuition here is that g is composed of two modes, but how to formalize this statement ? A mode is a local maximum of a function. Here, considering sufficiently small neighborhoods for the standard topology in $\mathbb{R}$, each point is a local maximum. For $r > 0$, the upper-level set $\mathcal{U}_r$ of g at level r is defined as $\mathcal{U}_r = \{i | g(X_i) \geq r\}$. An illustration is provided by figure 2.2.

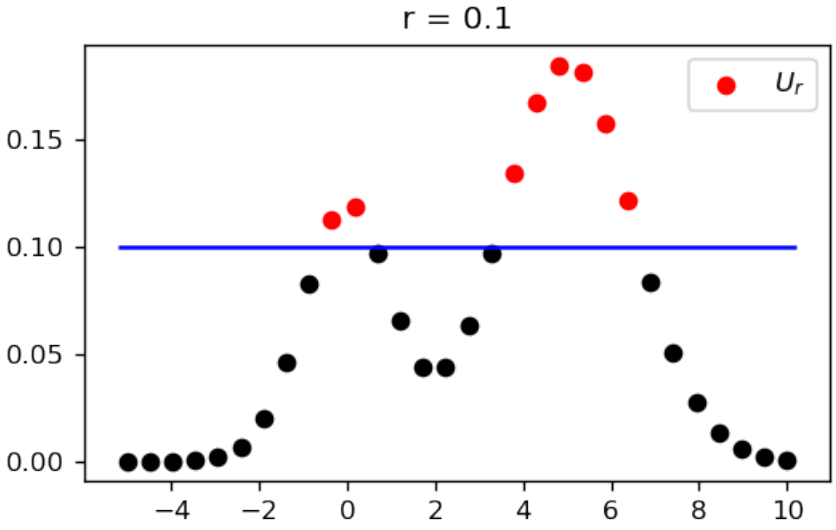


Figure 2.2: $\mathcal{U}_{0.1}$

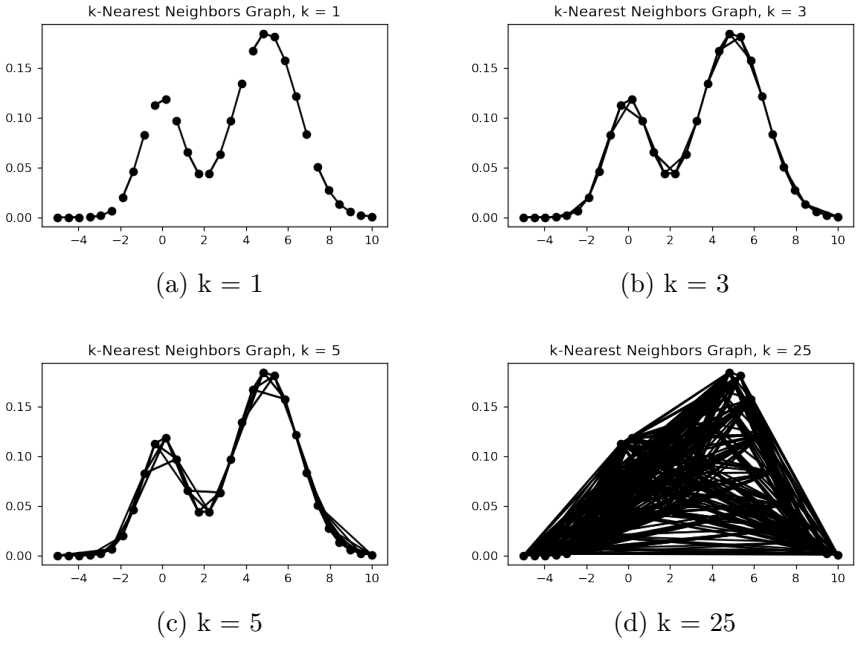
Intuitively, we can infer that $\mathcal{U}_{0.1}$ is composed of two disjoints connected components. However, from a topological point of view, it is just the union of isolated points.

k-Nearest Neighbors' Graph We need to be able to infer topological properties from discrete samples. It is precisely the aim of Topological Data Analysis (TDA). An introduction can be found in [13]. The general objective is to extract topological information based on discrete sets of points. To do so, it is necessary to build continuous objects over a point cloud. It can be done using neighboring graphs if the only topological information we need to recover is connectivity.

A neighboring graph is a graph whose vertices are the indices $\{1, \dots, N\}$,

and an edge between i and j means that X_i is in the neighbor of X_j and vice versa. Assume that we have a distance d on the sample $X_1, \ldots, X_n$, the most popular options are k-Nearest Neighbors (k-NN) graph, for k a positive integer or ϵ-proximity graph for $\epsilon > 0$. The k-NN graph is constructed with the rule: there is an edge between i and j if X_j is among the k nearest neighbors of j for the distance d or vice versa. The ϵ-proximity graph is constructed with the rule : there is an edge between i and j if $d(X_i, X_j) \leq \epsilon$. In the sequel, we will consider that d is the Euclidean distance.

The advantage of k-NN graph is that the parameter k is not hard to set with little knowledge on the underlying structure. Oppositely, choosing a value for ϵ necessitates to have a prior idea of the size of the neighborhoods. If the points belong to $\mathbb{R}^d$, the k-NN graph is robust to the dilatation of the sample. The main drawback of k-NN is that it can not isolate a point, even if it is very far from the rest of the sample. In the sequel, we will only consider k-NN graph. In the one-dimensional example, neighbors' graph can be easily visualized. The k-NN graph in our example for different values of k is showed in figure 2.3



(a) k = 1 (b) k = 3

(c) k = 5 (d) k = 25

Figure 2.3: k-NN graph for different values of k

The choices $k = 3$ and $k = 5$ seem consistent with an intuitive notion of

neighborhood in this example. $k = 1$ is a bad choice here because it creates too local neighborhood, leading to 8 connected components. $k = 25$ is also an undesirable choice here, as almost all points of the sample are neighbors of almost all other points. In general, we must select a value of k operating a compromise between a too small value leading to many undesirable connected components and too big values leading to a loss of information about proximity. For the illustrative example, we fix $k = 5$.

ToMATo relies on the possibility of implementing two operations: finding the modes of the function g and computing the topology of the upper-level sets of g. We will see how it is possible using a neighboring graph.

Mode seeking In the continuous setting, if we assume that the function g is twice differentiable, with a finite number of critical points and all its critical points are non-degenerate, the modes of g can be defined by the gradient flow. For $x \in \mathbb{R}^d$, the gradient ascent flow starting at x is the curve defined by the following differential equation defined for $t \geq 0$.

$$\begin{cases} \gamma_x(0) = x \\ \dot{\gamma}_x(t) = \nabla g(\gamma_x(t)) \end{cases} \tag{2.1.1}$$

Roughly speaking, it consists in starting at x and climb along the graph of g. If x is not a local minimum, $\gamma_x(t)$ tends to a local maximum of g as $t \to \infty$. Then, except for the union of a finite number of points (the local minima of g), every point in $\mathbb{R}^d$ is associated with a unique local maximum. Given a local maximum, $\bar{x}$, the set of all points for which the gradient ascent flow converge to $\bar{x}$ is called the mode associated with $\bar{x}$.

In a discrete setting, we have to find a proxy for the gradient ascent in order to identify the modes. The solution proposed by ToMATo is to create a graph whose vertices are the indices $\{1, \ldots, N\}$ and where for all $i \in \{1, \ldots, N\}$, i is connected with the index corresponding to the maximum value of g among the neighbors of X_i defined by the neighboring graph. The edges can be seen as steps in a discrete gradient ascent procedure. In the case where the neighboring graph is a k-NN graph, this graph is called the k-modes graph and its connected components are the k-modes of the sample. As illustrated in figure 2.4, in our example with the parameter $k = 5$ we obtain 2 modes.

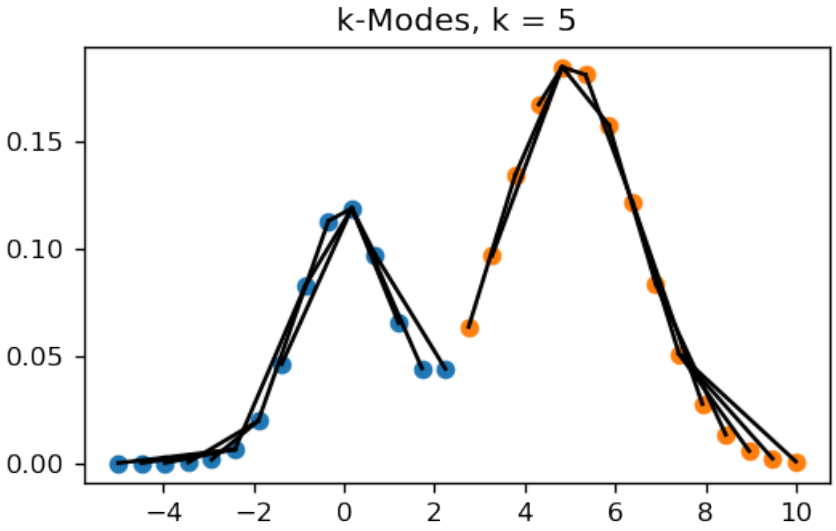
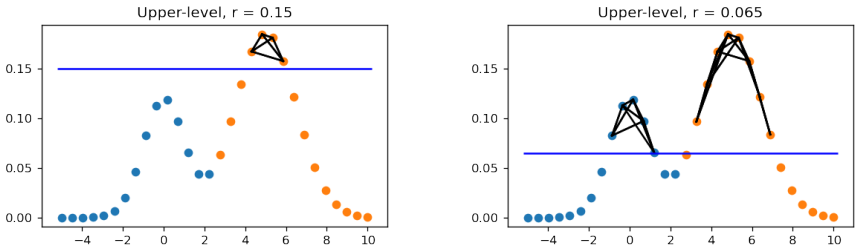


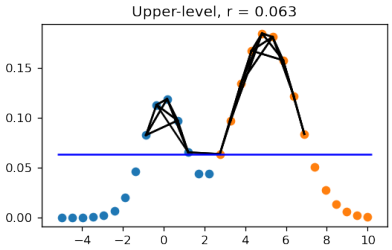
Figure 2.4: 5-Modes graph

Topology of upper-level and persistence diagram Now we are in position to describe the functioning of ToMATo. Given a real parameter r, we can define an upper-level set of g incorporating non-trivial topological information thanks to the k-NN graph. The upper-level $\mathcal{U}_r^k$ is the sub-graph of the k-NN graph for which all vertices i such that $g(X_i) < r$ were removed. The evolution of $\mathcal{U}_r^k$ when r goes from $+\infty$ to $-\infty$ determines a sequence of topological events. In our example, an illustration is given by figure 2.5.

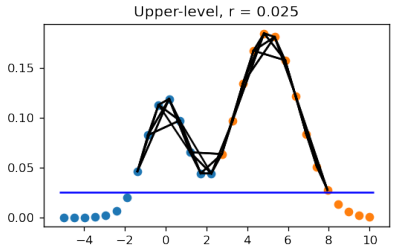


(a) r = 0.15

(b) r = 0.065



(c) r = 0.063



(d) r = 0.025

Figure 2.5: $\mathcal{U}_r^5$ for different values of r

During the evolution of r from $+\infty$ to $-\infty$, we observe four phases. If r is greater than the global maximum of g, then $\mathcal{U}_k^r = \emptyset$. When r lies between the maxima of blue and orange mode (figure 2.5a), $\mathcal{U}_k^r$ is composed of one connected component. Then in the situation illustrated by figure 2.5b, $\mathcal{U}_k^r$ is composed of two connected components. After that, the two connected components get merged, as illustrated in figures 2.5c and 2.5d.

The transitions between these phases correspond to topological events, which can be the appearance (birth) or the merge (death) of a connected component into another. When a merge occurs, we consider that the connected components corresponding to the mode with the lowest density dies and get merged into the connected component associated with the mode of highest density. These topological events can only occur when $r = g(X_i)$ for some $i \in \{1, \ldots, N\}$, making the record of all topological events implementable. Note that at least one mode is never merged, so its death time is $-\infty$

Now, each mode is characterized by a couple (birth-time, death-time), this information can be summarized in a Persistence Diagram (PD) [18]. Note that

our definition slightly differs from the usual one as in our example birth-times are higher than death time, but the associated results are the same.

Definition 2.1.1

A persistence diagram is a multi-set of points of $\mathbb{R}^2$ located under the diagonal. A multi-set is a collection of points $(x_i, y_i)_{1 \leq i \leq k}$ where y_i can be equal to $-\infty$ associated with multiplicities $(m_i)_{1 \leq i \leq k} \in \mathbb{N}_*^k$

In our illustrative example, the PD is given by figure 2.6, the colors of the points corresponds to the colors of the modes in figure 2.4, and the multiplicity associated to both points is one. The PD is a tool to summarize topological information about data.

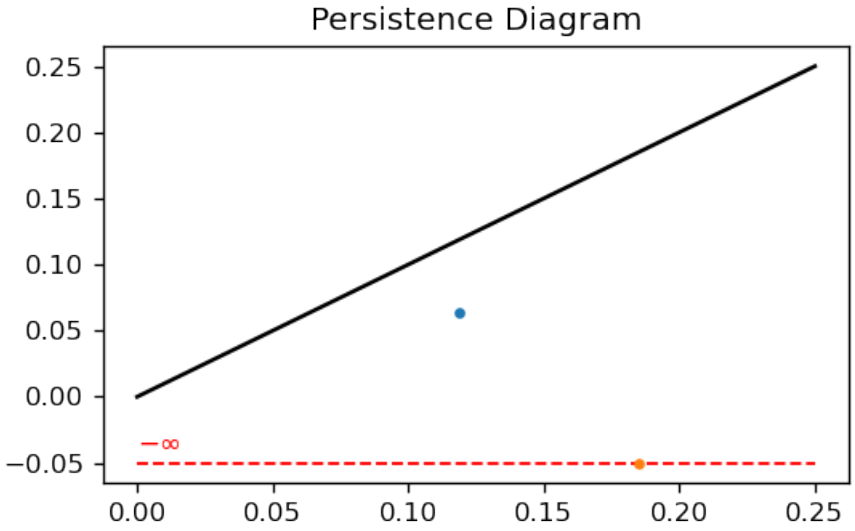
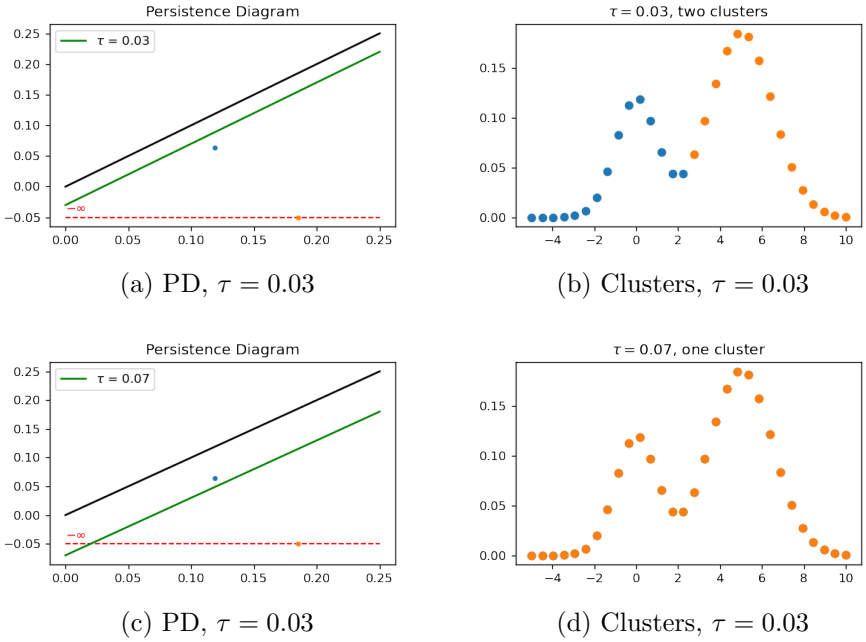


Figure 2.6: The PD for our example, the red dashed line correspond to an ordinate of $-\infty$

Prominence The difference between death-time and birth-time for a given mode is called its prominence. This quantity summarizes an intuitive notion of importance of each mode.

Clustering Now, we set a positive threshold parameter τ. To obtain a clustering of the initial data, we start with one cluster per modes. Then we merge clusters following the possible merges indicated by the procedure illustrated in figure 2.5 with the rule that only modes of prominence lower than τ are merged. In the illustrative example, two situations are possible, depending on whether the threshold parameter τ is higher or lower than the prominence level of the blue mode. These two configurations are illustrated by figure 2.7.



(a) PD, $\tau = 0.03$ (b) Clusters, $\tau = 0.03$

(c) PD, $\tau = 0.03$ (d) Clusters, $\tau = 0.03$

Figure 2.7: Clustering output for different values of τ

ToMATo The algorithm which takes as inputs a collection of points $X_1, \ldots, X_N$, equipped with a distance d, a function g over the data, a neighboring parameter k, a positive threshold τ and outputs a clustering of the data is called ToMATo and were introduced in [14]. For the sake of clarity, we have presented the algorithm in three steps: mode identification, construction of PD with upper-level sets topology and merging step. From a computational point of view, it is possible to combine all these steps in one by updating a union-find structure in a loop over data points by decreasing g values. The algorithmic cost of the operation is $N \log N$, the cost of sorting the data by g values.

2.1.2 The stability theorem

The advantage of ToMATo as density-based clustering algorithm is that it is robust to small perturbation of the input function g. This property is a consequence of a more general result of TDA known as stability theorem ([18], [15], [16]) stating that the PD representation enjoys a Lipschitz property.

Let us consider an illustrative situation where the data $X_1, \ldots, X_N$ is the same as previously but the function g is replaced by $\tilde{g}$ which if a random perturbation of g: for all $1 \leq i \leq N$, $\tilde{g}(X_i) = g(X_i) + \epsilon_i$ where $\epsilon_1, \ldots, \epsilon_N$ are iid variables uniformly distributed on $[-0.02, +0.02]$. This situation is represented in figure 2.8

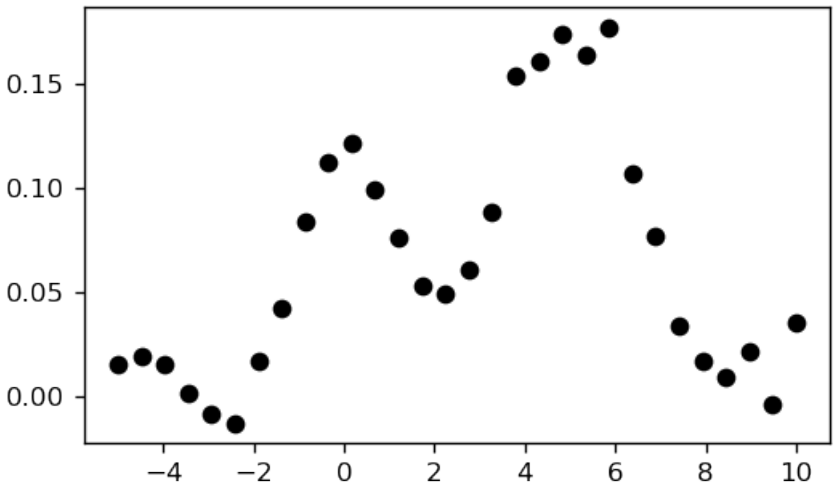
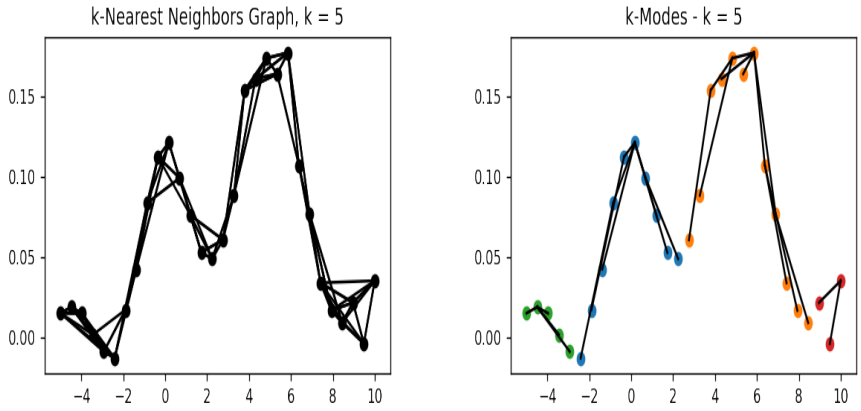


Figure 2.8: Perturbed dataset

Applying the same methodology as previously with $k = 5$, we obtain k-NN graph and modes as illustrated in figure 2.9.



(a) 5-NN graph (b) 5-Modes

Figure 2.9: 5-NN graph and 5-Modes for perturbed dataset

The k-NN graph does not change, as it only depends on the pairwise distances between the sample points. Oppositely, the mode seeking phase depends on the values of the function. Here, for $\tilde{g}$, four modes were found instead of two for g. We can now compute the PD corresponding to this situation.

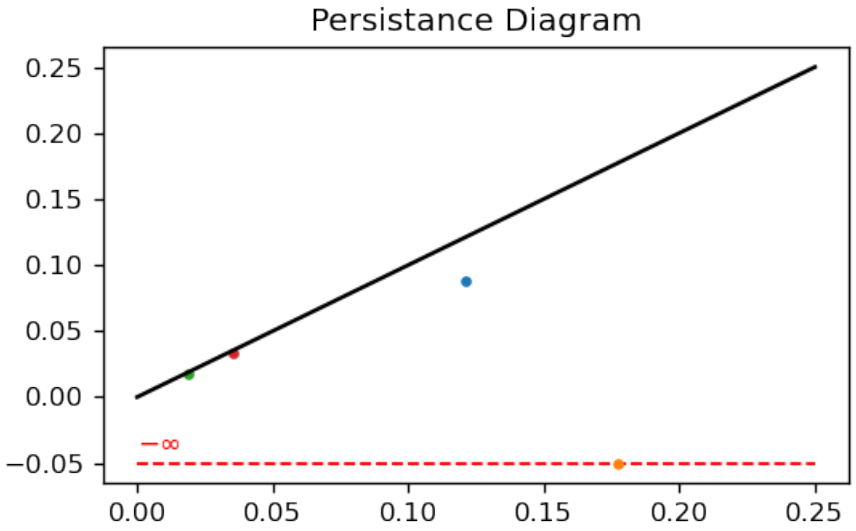
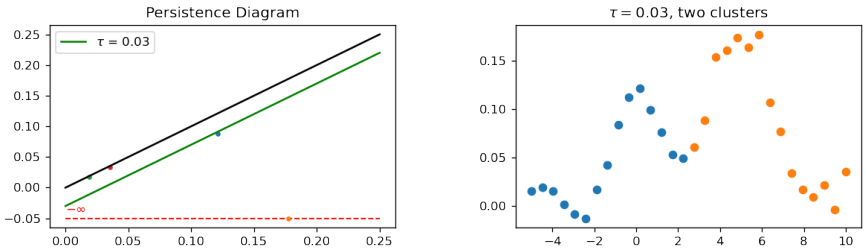


Figure 2.10: PD for perturbed dataset

We observe that the result looks very similar to what was obtained in figure 2.6, with the difference that we have two extra modes (green and red) characterized by a small prominence value (their representations on the diagram are very close to the diagonal).



(a) PD, $\tau = 0.03$ (b) Clusters, $\tau = 0.03$

Figure 2.11: Clustering output for different values of τ

Bottleneck distance We have seen that PDs are tools to topologically describe a function over a point cloud via the notion of prominence. A natural question to ask is the possibility of defining a distance between diagrams, with in mind that close diagrams mean close functions in some sense. The bottleneck distance is a popular distance between diagrams.

Definition 2.1.2

Let us consider two persistence diagrams, PD_1 and PD_2. A collection of couples of points $((x_1, y_1), (x_2, y_2))$ where (x_1, y_1) belongs to PD_1 or to the diagonal $(x_1 = y_1)$ and where (x_2, y_2) belongs to PD_2 or to the diagonal is a matching if every point of PD_1 and PD_2 are present in γ with their respective multiplicity. The possibility of matching points with the diagonal allows defining matching between diagrams with possibly different number of points.
We define the sup-norm of a matching γ by

$$\|\gamma\|_\infty = \sup_{((x_1,y_1),(x_2,y_2))\in\gamma} \max\left\{|x_1 - x_2|, |y_1 - y_2|\right\}. \tag{2.1.2}$$

The bottleneck distance d_B between PD_1 and PD_2 is the sup-norm of the matching between PD_1 and PD_2 with the lowest sup-norm. Formally

$$d_B(PD_1, PD_2) = \inf_\gamma \|\gamma\|_\infty \tag{2.1.3}$$

where the inf is taken over all possible matching.

In our example, we can compute the bottleneck distance between $PD(g)$ and $PD(\tilde{g})$, the persistence diagrams associated respectively to the functions g and $\tilde{g}$. The optimal matching is represented in figure 2.12. The two points of $PD(\tilde{g})$ near the diagonal are matched to their projections on the diagonal.

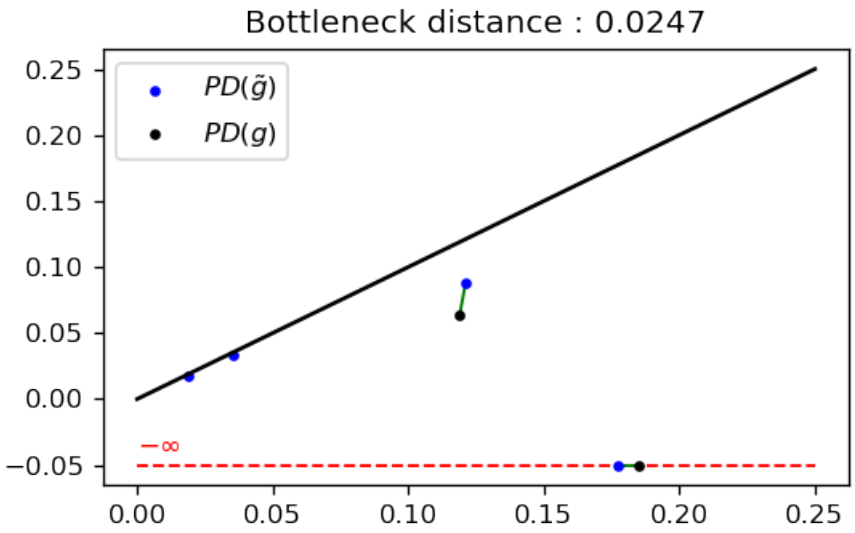


Figure 2.12: Bottleneck distance between PDs

Theorem 2.1.3: Consequence of theorem 4.4 in [15]

Let g and $\tilde{f}$ two functions over $X_1, \ldots, X_N$ and consider the associated persistence diagrams PD_k and $\tilde{PD}_k$. We have that

$$d_B(PD_k, \tilde{PD}_k) \leq \|f - \tilde{f}\|_\infty \tag{2.1.4}$$

This result indicates the stability of the persistence diagram representation with respect to the sup norm for the functions. It is interesting to have such a result as in practice, the density g over the data points is not know but must be estimated.

2.1.3 Hierarchical clustering with ToMATo

An interesting fact about ToMATo is that it is possible to extract a hierarchical clustering by moving the value of the threshold parameter τ from $+\infty$ to 0. Note that this hierarchical structure computation does not have a greater algorithmic complexity than ToMATo with a given threshold, as all the information about

prominence is stored. This structure can be represented by a binary tree, where each leaf corresponds to a mode and node represent the merging of both children.

This is of particular interest for the application to cytometry data. Delivering a hierarchical structure rather than a flat clustering corresponding to a fixed value of τ allows the cytometrists to explore the clustering structure simultaneously at different level of granularity, which appears necessary knowing that cellular populations are often very unbalanced.

From a computational point of view, a hierarchical structure is the union of two data structures

- A clustering $C_1, \ldots, C_M$ corresponding to the modes (or a clustering for $\tau = 0$). If we think of a tree representation, it corresponds to the leafs.
- An collection of triplets (i, j, τ). (i, j, τ) is in the collection if clusters C_i and C_j merges at level τ to form a new cluster. The convention for the naming of the cluster is that C_{M+1} corresponds to the merging of two clusters in $C_1, \ldots, C_M$ with the lowest value of τ, C_{M+2} corresponds to the merging of two clusters in $C_1, \ldots, C_M, C_{M+1}$ with the second-lowest value of τ and so on. All triplets are represented by a node in a tree representation.

2.2 DTM-based density estimation

ToMATo takes as input a function over the point cloud $X_1, \ldots, X_N$. In the context of clustering, this function must be thought as a density over the point cloud. Several approaches exist to estimate a density. After some quick comparisons with Kernel Density Estimators (KDE) on cytometry data, we chose to focus on Distance-to-Measure (DTM) based density estimators [6] because we found it harder to set the hyper-parameters of KDE in this context.

The DTM density estimator relies on the distance from each point of the sample to their neighbors. Here by distance we mean Euclidean distance in $\mathbb{R}^d$. The definition relies on the choice of a hyper-parameter k_d

$$g_{k_d}^{\mathrm{DTM}}(X_i) = \frac{k_d(1+k_d)}{2NV_d \sum_{j=1}^{k_d} \|X_i - X_{i(j)}\|^d} \tag{2.2.1}$$

where $X_{i(j)}$ is the j-th nearest neighbor of X_i among the points of the sample and V_d the volume of the unit ball in $\mathbb{R}^d$.

An alternative to $g_{k_d}^{\mathrm{DTM}}$ is given by the log-DTM density estimator, $g_{k_d}^{\mathrm{logDTM}} = \log\left(g_{k_d}^{\mathrm{DTM}}\right)$. The heuristic behind the use of log-density is that this will lower the relative difference of highness between the peaks of the density. It seems relevant in the context of cytometry, where populations are often unbalanced, leading to high differences in density level. Applying a logarithm might have the effect to temper such differences and equalize the prominence of different size cell populations.

2.3 CyToMATo

We are now in position to present our strategy to cluster cytometry data. The first step is to apply a DTM-based density estimation on $X_1, \ldots, X_N$ and then execute ToMATo. Three hyper-parameters have to be set: k_d and k_t, the neighboring parameters for density estimation and for ToMATo respectively, and decide if we use a DTM or log-DTM density estimator.

As we target high dimension applications, our reference benchmark will be the one of Weber and Robinson [85], with data of dimension 13 to 39. However, it is unfair to use the same data to calibrate the hyper-parameters and to evaluate the performance of the clustering. This is why we rely on the data used in the FlowCAP-I challenge [74] to set the hyper-parameters.

2.3.1 Evaluation setting

Data For the FlowCAP-I challenge, five cytometry experiments were analyzed. Four come from human blood with different pathologies and one from mouse blood with human hematopoietic stem cell transplant. Each of these cytometry exper-

iments was replicated between 10 and 30 times. We used a subsample of three replication per experiments with available labels to evaluate our method.

- GvHD (Graft-versus-Host Disease) with $N \simeq 15,000$ and $d = 6$
- DLBCL (Diffuse Large B-cell Lymphoma) with $N \simeq 5,000$ and $d = 5$
- ND (Normal Donors) with $N \simeq 60,000$ and $d = 12$
- WNV (symptomatic West Nile Virus) with $N \simeq 90,000$ and $d = 8$
- HSCT (Hematopoietic Stem Cell Transplant) with $N \simeq 9,000$ and $d = 6$

For the Weber and Robinson's benchmark, four datasets were analyzed. Two are experiments on human bone marrow cells from healthy donors and to on mouse bone marrow cells.

- Levine13 (bone marrow cells from one healthy donor) with $N = 81,747$ and $d = 13$
- Levine32 (bone marrow cells from two healthy donors) with $N = 104,184$ and $d = 32$
- Samusik01 (bone marrow cells from mice) with $N = 53,173$ and $d = 39$
- SamusikAll (bone marrow cells from mice) with $N = 514,386$ and $d = 39$

Evaluation method The comparison between clusters and labeled populations are done using the F1 score. Let consider that we have two partitioning of the indices $1, \ldots, N$. $\mathcal{C} = (C_1, \ldots, C_K)$ is a set of clusters, obtained by any clustering algorithm and $\mathcal{P} = (P_1, \ldots, P_{K'})$ a set of populations, labeled by an expert. We aim to quantify the closeness between $\mathcal{C}$ and $\mathcal{P}$. A first step is to measure the correspondence between a cluster C_i and a population P_j through the F1 score.

Definition 2.3.1

Let C and P be two sets of indices in $\{1, \dots, N\}\}$. The precision of C for the identification of P is given by

$$P(P, C) = \frac{|P \cap C|}{|C|}. \tag{2.3.1}$$

The recall C for the identification of P is given by

$$R(P, C) = \frac{|P \cap C|}{|P|}. \tag{2.3.2}$$

The precision measure the proportion of indices of C belonging to P and the recall the proportion of indices of P belonging to C.
The F1 score between C and P is the harmonic mean between the precision and the recall

$$F1(P, C) = \frac{2 \times P(P, C) \times R(P, C)}{P(P, C) + R(P, C)}. \tag{2.3.3}$$

Its value lies between 0 and 1. A F1 score of 0 means that $P \cap C = \emptyset$, a F1 score of 1 means that $P = C$.

The collection of F1 scores $F1(P_j, C_i)_{1 \leq j \leq K', 1 \leq i \leq K}$ can be stored in a $K' \times K$ table. Then we can associate each population to a cluster in order to maximize the sum fo the F1 scores corresponding to the associations. If $K' > K$, meaning that there are more populations than clusters, some populations will not be associated to any cluster and the corresponding F1 score will be zero. In practice, the association is computed with the Hungarian algorithm [40]. At this step, we obtain a list of F1 score, one for each labeled population, $F1(P_1, \mathcal{C}), \dots, F1(P_{K'}, \mathcal{C})$.

From $F1(P_1, \mathcal{C}), \dots, F1(P_{K'}, \mathcal{C})$, we want to define a single value as a performance quantification of a given algorithm. The first proposition is the unbalanced global F1 score, used in the FlowCAP-I challenge. The definition of this global criterion is given by

$$F1_{\mathrm{U}}(\mathcal{P}, \mathcal{C}) = \sum_{j=1}^{K'} \frac{|P_i|}{N} F1(P_i, \mathcal{C}) \tag{2.3.4}$$

The issue with that criterion is that it give more importance in the detection of big populations. It could be argued that in the context of cytometry it is unfair,

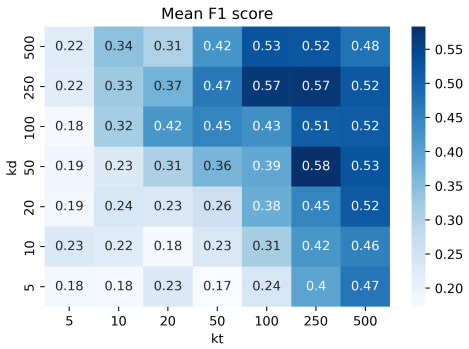
as the biological importance of a population is not related to its size. This is why this criterion was replaced in more recent benchmarks, as the one of Weber and Robinson by the balanced global $F1$ score

$$F1_{\mathrm{B}}(\mathcal{P}, \mathcal{C}) = \frac{1}{K'} \sum_{j=1}^{K'} F1(P_i, \mathcal{C}). \tag{2.3.5}$$

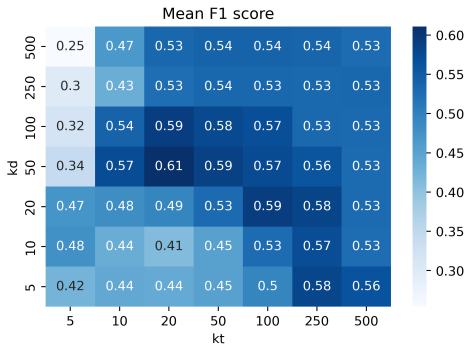
2.3.2 Hyper-parameters calibration

In order to study the impact of the choice of the parameters on the quality of the clustering for DTM and log-DTM density estimators, we ran ToMATo with multiple combination of these for the datasets proposed by FlowCAP-I. The number of clusters outputted by ToMATo is set to be the true number of populations, we do not evaluate the capability of our algorithm to output the exact number of clusters as we focus on the hierarchical structure as an output.

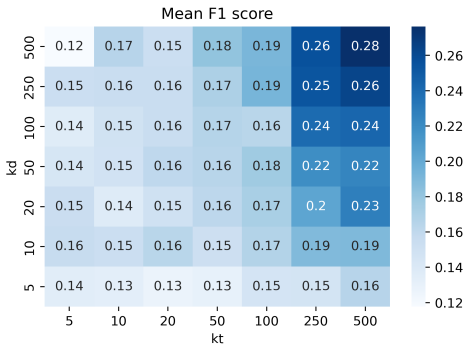
Results for DTM density estimators are presented in figure 2.13 and for log-DTM density estimators in figure 2.14. The indicated values are the mean value of the mean F1 score across populations for the three replications of all experiments.



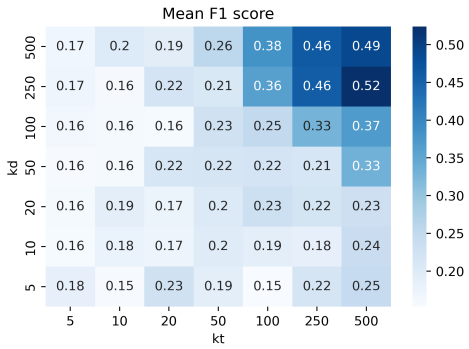
(a) GvHD



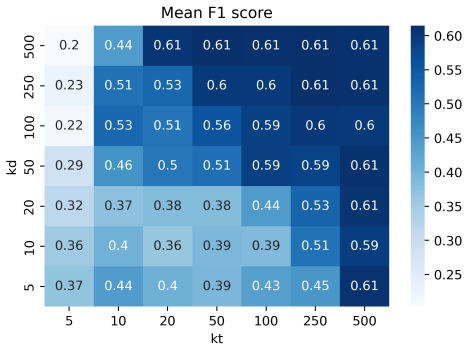
(b) DLBCL



(c) ND

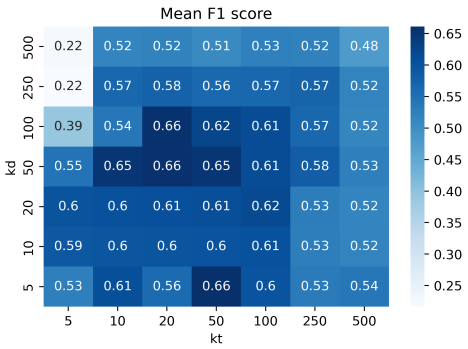


(d) WNV

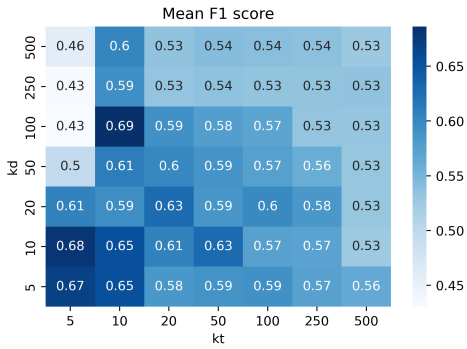


(e) HSCT

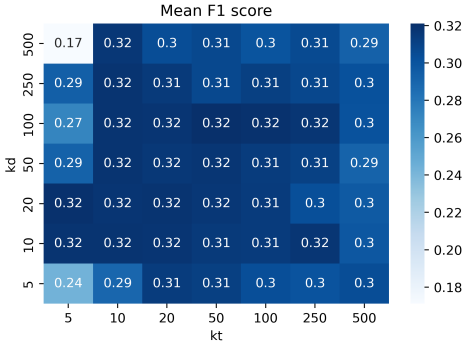
Figure 2.13: Mean F1 scores for ToMATo with DTM density with respect to k_d and k_t on the datasets from FlowCAP-I challenge



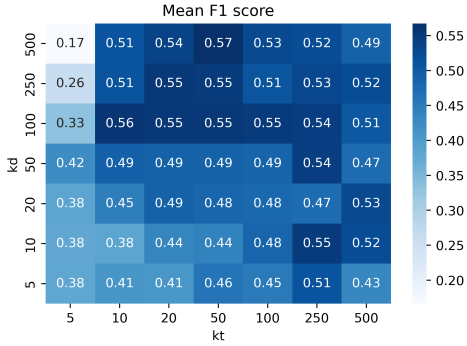
(a) GvHD



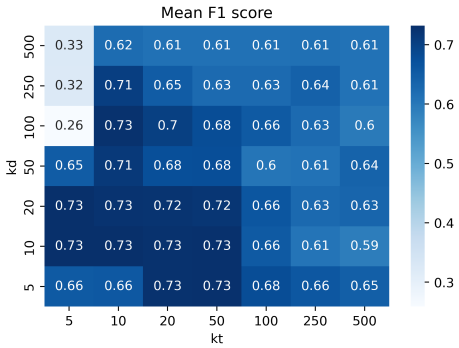
(b) DLBCL



(c) ND



(d) WNV



(e) HSCT

Figure 2.14: Mean F1 scores for ToMATo with log-DTM density with respect to k_d and k_t on the datasets from FlowCAP-I challenge

We remark that when both k_d and k_t take high values (greater than 100) the performances of ToMATo based on DTM density and log-DTM density are comparable. At the opposite, for small values of both parameters (50 or less), the usage of log density significantly improve the quality of the clustering. In general, log-DTM density seems to be the best choice, we select it as the standard choice for analyzing cytometry data with ToMATo.

Now, regarding k_d and k_t, it is interesting to observe a great stability of performance for values of both parameters lying between 5 and 50 across all datasets. This is of great interest to us as it indicates that their parameters can be safely hard-coded. $k_d = k_t = 20$ seems to be a reasonable choice. For these reasons, we have decided that it could be interesting to hard-code the values of k_d and k_t at a value of 20 and use a log-DTM density estimator. The advantage is that our method is now hyper-parameter-free.

2.3.3 The algorithm

We have just seen that, based on the results for the FlowCAP-I data, we can safely hard-code the parameters k_d and k_t and fix the density estimation step to the logarithm of a DTM-based density estimation step over $X_1, \ldots, X_N$. This results to the algorithm CyToMATo (algorithm 1).

input : $X_1, \ldots, X_N \in \mathbb{R}^d$
output: A hierarchical structure
begin
  Compute the 20-nearest neighbors graph of $X_1, \ldots, X_N$ and store the corresponding distances.
  Compute $g_{k_d}^{\mathrm{logDTM}}(X_i)$ for $i = 1, \ldots, N$
  Compute the 20-NN graph
  Run ToMATo
end

Algorithm 1: CyToMATo

2.3.4 Validation on high-dimensional data

To measure how CyToMATo performs comparatively to other clustering algorithms for cytometry data, we compare the mean F1 scores obtained in [85] to the ones obtained with our approach. The number of outputted clusters is set to 40 as in the protocol described in the paper. Results are presented in figure 2.15. Methods are sorted from the one with the higher average for mean F1 score over the four dataset (FlowSOM, on top) to the one with the lower one (SWIFT). Regarding the ranking and the relative values of F1 scores, it is fair to say that CyToMATo is comparable to the state-of-the-art algorithms for the problem of recovering cellular populations based on cytometry measurements. What is more, it has the great advantage to be hyper-parameter-free.

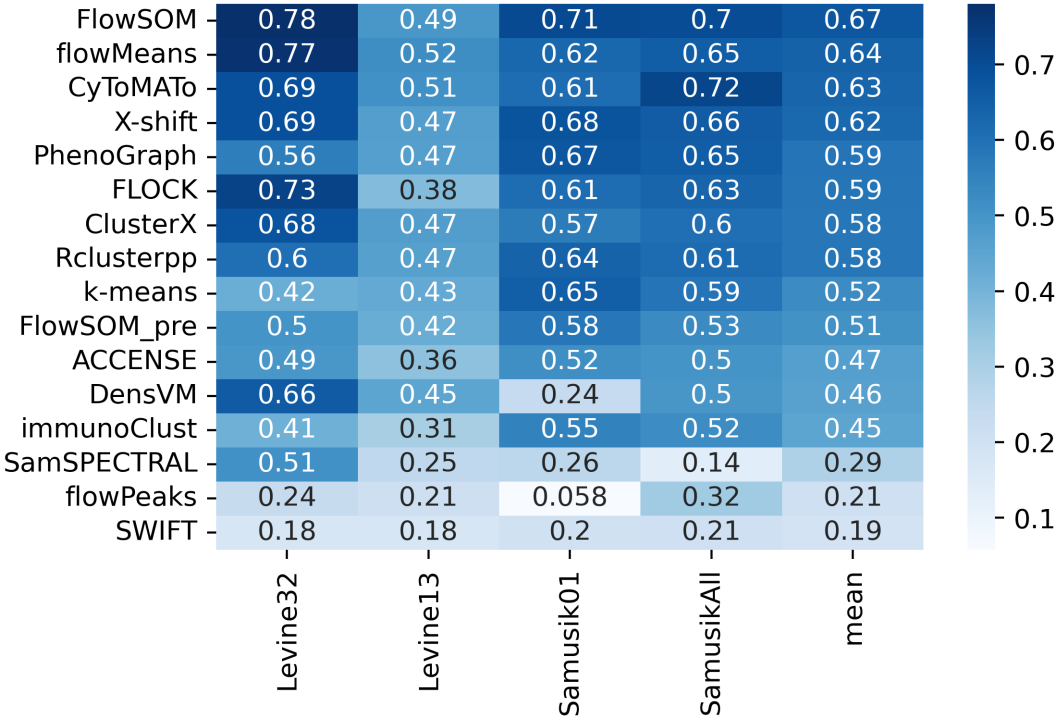


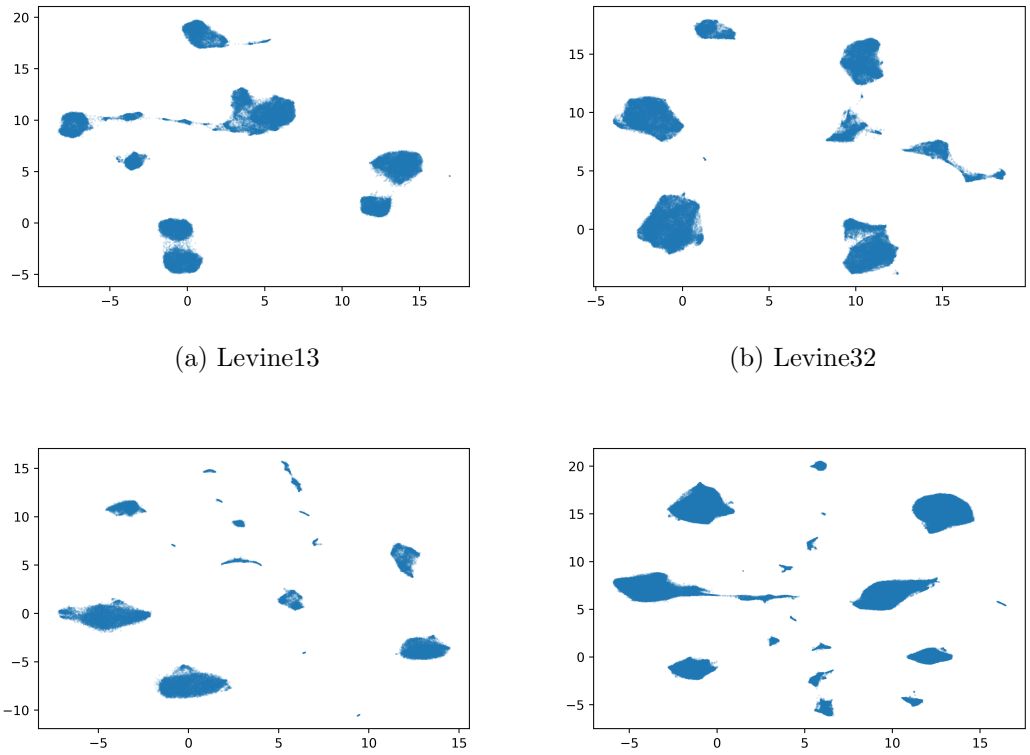
Figure 2.15: Comparison of clustering scores of ToMATo and other algorithms tested in [85] for Levine13, Levine32, Samusik01 and SamusikAll and the mean of the four scores

This study was intended to give an idea of how CyToMATo could be a

relevant tool to cluster cytometry data. The choice of the benchmarking method is questionable. Is it of interest because it allows to compare easily between different algorithms, but it misses to give full account of the specificity of each algorithm to efficiently identify cell populations. At this point, we identified two paths for improvement. Firstly, CyTOMATo does not include any dimensionality reduction idea. This is clearly something we have to working on, as the impact of the dimensionality of the data is one of the main challenge today for cytometry data analysis. Secondly, we aim to be able to give more than just a clustering but qualitative information about the data that could be useful for the cytometrist.

2.4 Dimensionality reduction via UMAP

Probably the most popular approach for dimensionality reduction in the field of cytometry, UMAP (Uniform Manifold Approximation and Projection) aims to propose a low-dimensional embedding of a high dimension point cloud. The algorithm was proposed first in [51]. Specific applications to cytometry data are detailed in [86] or [5] for example. A recent benchmark study [83] compares UMAP with alternative algorithms for dimensionality reduction and studies the impact of the hyper-parameters. The aim of UMAP is to propose an embedding $Y_1, \ldots, Y_N$ of the dataset $X_1, \ldots, X_N$. Typically, Y_i belongs to $\mathbb{R}^2$ or $\mathbb{R}^3$. The publication of UMAP were accompanied by the publication of a python package. Using the default parameters, we have applied UMAP to the data of the Weber and Robinson's benchmark. Low-dimensional embeddings can be visualized in figure 2.16. In these representations, we clearly see that a structure in clusters is obtained in the representation in $\mathbb{R}^2$.



(a) Levine13

(b) Levine32

(c) Samusik01

(d) SamusikAll

Figure 2.16: UMAP projections

As an alternative to CyToMATo on $X_1, \ldots, X_N$, one can run CyToMATo on the low-dimensional data outputted by UMAP, following again the protocol of [85]. The results are presented in figure 2.17.

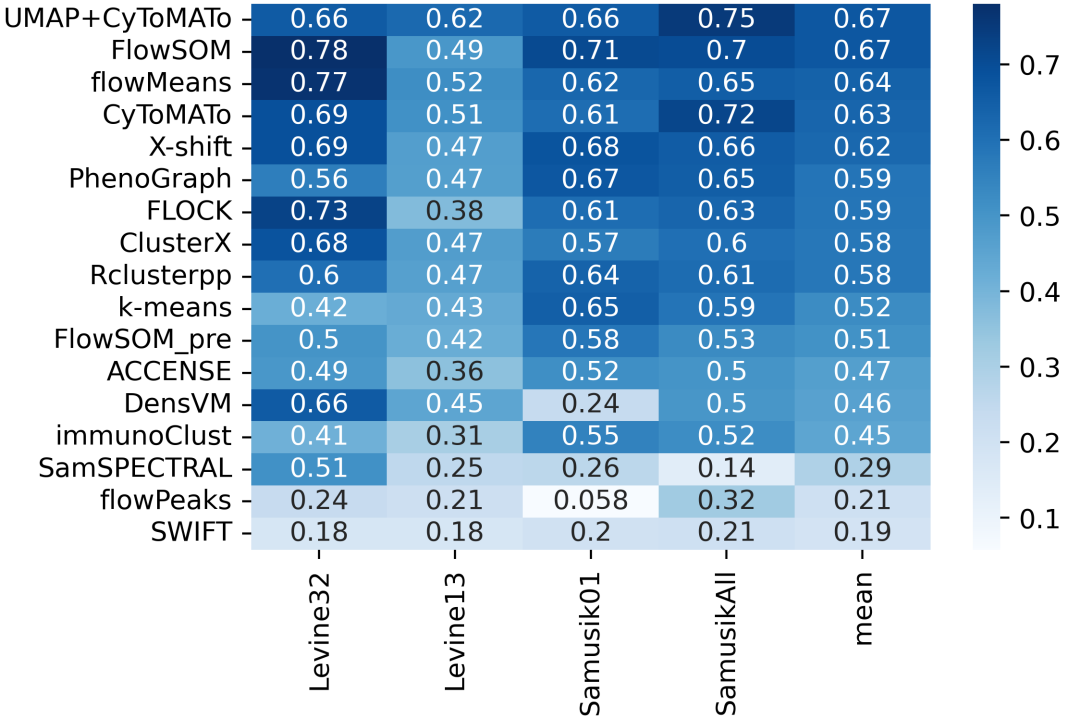


Figure 2.17: Comparison of clustering scores of CyTOMATo, and UMAP+CyToMATo and other algorithms tested in [85] for Levine13, Levine32, Samusik01 and SamusikAll and the mean of the four scores

The conclusion is that a dimensionality reduction step has the potential to increase the performance of CyToMATo. However, the functioning of UMAP is quite opaque, and the embedding is not easily interpretable. What is more, it is sensitive to the initialization, this fact was highlighted in the context of cytometry in a recent paper [37]. So, UMAP has to be thought as a powerful tool for exploratory experiments, but it should be more proficient to find more explainable and robust tools to target clinical applications.

2.5 Conclusion

In this chapter, we have developed an approach for cytometry data clustering based on the combination of DTM-based density estimation and deterministic hierarchical structure detection with ToMATo. What is more, we have shown on the data provided by the FlowCAP-I challenge [1] that using log densities, the outputs were not too much sensitive to the choice of the hyper-parameters. Then, we have defined CyToMATo with fixed hyper-parameter. The efficiency of the method have been tested on the data provided by Weber and Robinson [85]. In comparison with existing approaches, our method is comparable to state-of-the-art algorithms in terms of F1 score.

What is more, we have shown the potential of improvement with the addition of a preprocessing step in the procedure, a dimensionality reduction step with the algorithm UMAP. This has led to better clustering performances. However, UMAP acts as a black-box method. In the rest of this work, we will develop tools to incorporate dimensionality reduction in the density estimation step through an interpretable model: the independence structure.

Chapter 3

Independance Structure Density Estimation: a computationally efficient approach for density estimation under Independance Structure model

Context The model of Independence Structure (IS) was introduced by Lepski [42] and its minimax risk was studied by Rebelles [65] under L_p loss, $1 \leq p < \infty$ and Lepski [42] for the L_∞ loss. Conceptually, their results indicate that if the IS is met for the unknown density, the complexity of the density estimation task is no longer related to the ambient dimension but rather to the size of the biggest block in the decomposition of the variables into independent blocks. The IS is also of interest for us as it is easily interpretable and seems to be relevant in the context of cytometry. However, no algorithmic approach has been developed to learn en IS from the data with reasonable data sizes. The estimators introduced in [42] and [65] rely on brute-force approaches to select the best partition of variables, which is too costly when the dimension grows.

Our contribution We present Independence Structure Density Estimation (ISDE), a method designed to simultaneously compute a partition of the features and a density estimation as a product of marginals over this partition in order to maximize the empirical log-likelihood, or equivalently, minimize the KL loss. This change of objective function is not purely decorative: it allows reducing the algorithmic complexity by reducing the size of manipulated data structure from the number of partitions of the variables to the number of subsets of the variables. Our method enjoys reasonable running time for moderately high-dimensional problems and can be combined with any density estimation technique, so it covers parametric and nonparametric settings. To our knowledge, this problem were only addressed in the Gaussian case [22], we are the first to design an algorithm estimating as IS in the context of Kernel Density Estimation. Furthermore, our implementation relies on the usage of GPU (Graphical Processing Unit), which allows computational speed-up in comparison to standard approaches.

Organization of the chapter In section 3.1 we give a brief review of nonparametric density estimation aspects that will be of interest for us. In section 3.2 we present in detail the construction of ISDE. In section 3.3 we empirically prove that ISDE is efficient on simulated data, and in section 3.4 we analyze the algorithmic complexity and the running time of ISDE.

3.1 Nonparametric density estimation

Density estimation is a central topic in statistics. In this thesis, we focus on the iid setting: we suppose that the observation $X_1, \dots, X_N$ are iid realizations of some random variable over $\mathbb{R}^d$ admitting a density f with respect to the Lebesgue measure. The question density estimation tries to answer is: how accurately f can be estimated thanks from $X_1, \dots, X_N$. This question must be precised in order to derive mathematical results. In this section, we introduce some important notions in the field of density estimation to understand the problems and approaches.

Nonparametric and Parametric Density Estimation The easiest way to do density estimation is to consider parametric models: data is supposed to be drawn from a probability distribution known up to a finite-dimensional parameter θ. Estimating the density is then equivalent to estimating θ. One example is the centered multivariate Gaussian framework, where the parameter θ is the covariance matrix Σ. An introduction to parametric statistics can be found in [84], chapter 9. This approach suffers from a lack of flexibility, as it strongly constrains the model. At the other end of the spectrum lies nonparametric density estimation. In this framework, densities are no longer considered members of some finite-dimensional family but are supposed to belong to a set of functions with a given regularity. An introduction to the subject can be found in [77]. In this thesis, we will focus more on nonparametric density estimation.

Minimax Risk The minimax risk is a common measure of the complexity of a density estimation task. Let D be a (pseudo)distance on the space of density functions. If the true density is f and $\hat{f}$ is an estimator (a density function measurable with respect to the sample $X_1, \dots, X_N$), the risk of the estimator $\hat{f}$ is defined as

$$\mathcal{R}(D, f) := \mathbb{E}\left[D(f, \hat{f})\right]. \tag{3.1.1}$$

It correspond to the expected distance between f and $\hat{f}$ where the expectation is taken over the realizations $X_1, \dots, X_N$.

Now, we assume that the true density belongs to some known collection of

density functions $\mathcal{F}$, called a model, the minimax risk is defined as follows

$$\mathcal{R}(D, \mathcal{F}) := \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}\left[D(f, \hat{f})\right] \tag{3.1.2}$$

where the inf is taken over all measurable functions from the data to $\mathcal{F}$. More specifically, a great part of the literature on minimax risks deals with the asymptotic regime of $\mathcal{R}(D, \mathcal{F})$ with respect to N.

An example of model: Hölder Balls Let $\mathcal{U}$ be an open subset of $\mathbb{R}^d$ and $g : \mathcal{U} \to \mathbb{R}$ a function. Let $\gamma = (\gamma_1, \ldots, \gamma_d) \in \mathbb{N}^d$ be a multi-index and let $|\gamma| = \sum_{i=1}^d \gamma_i$ be its order. The partial differentiate operator D^γ is defined as follows

$$D^\gamma g = \frac{\partial^{|\gamma|} g}{\partial_1^{\gamma_1} \ldots \partial_d^{\gamma_d}}. \tag{3.1.3}$$

For a positive number β, if we denote by s the larger integer strictly lower than β and let $\delta = \beta - s \in (0, 1]$, g belongs to the Hölder ball $\mathcal{H}(\beta, L)$ where L is a positive real number if both following conditions are fulfilled

$$\begin{cases} \max\limits_{|\gamma| \leq s} \sup\limits_{x \in \mathcal{U}} \|D^\gamma g(x)\| \leq L \\ \max\limits_{|\gamma| = s} \sup\limits_{x, y \in \mathcal{U}} |D^\gamma g(x) - D^\gamma g_i(y)| \leq L \|x - y\|^\delta. \end{cases} \tag{3.1.4}$$

If g is defined as a close subset $\mathcal{C}$ of $\mathbb{R}^d$, we say that $g \in \mathcal{H}(\beta, L)$ if the restriction of g to the interior of $\mathcal{C}$ belongs to $\mathcal{H}(\beta, L)$.

Minimax Risk over Hölder Balls In [33], the minimax rate of this family of functions was studied considering L_p losses. In particular, the result with the squared L_2 distance is the following

$$\mathcal{R}\left(\|.\|_2^2, \mathcal{H}^\beta(d, H)\right) \asymp N^{-\frac{2\beta}{2\beta + d}}. \tag{3.1.5}$$

The minimax risk goes to zero when the number of observations goes to infinity. The speed of convergence is governed by the exponent $-\frac{2\beta}{2\beta+d}$. We can

interpret this bound as a manifestation of the curse of dimensionality because of its dependence on the dimension d. For practitioners, it should be adventurous to use a multivariate density estimator if the sample size is limited and the dimension becomes large, especially in the case of nonparametric estimation. A solution is to assume that unknown density belongs to a class of structured functions. A solution to obtain a faster rate of convergence is to add some structural conditions to the model.

Structural Density Estimation with Undirected Graphical Models A way to consider a structure for a multivariate random variable is through its undirected graphical model (introduction to the field can be found in [27] and more in-depth cover in [82]). As we will not consider directed graphical models, we always consider that graphs are undirected in the sequel. Given a graph $G = (V, E)$ whose vertices correspond to the features $\{1, \dots, d\}$ we say that G is a graphical model for X if the following condition is satisfied:

$$(i, j) \notin E \Rightarrow X^i \perp\!\!\!\perp X^j | (X^k)_{k \notin (i,j)}. \tag{3.1.6}$$

Constraints on the graphical model associated with a distribution impose a structure on the density, and such a structure can help overcome the curse of dimensionality. However, learning a graphical model is a complex task in many situations. The general result is that if G is a graphical model for a d-dimensional random variable X, denoting by $\mathcal{C}$ the set of cliques of G (*ie* fully connected sets of nodes), it exists a collection of nonnegative functions $(\psi_C)_{C \in \mathcal{C}}$ such that the density f of X can be written as

$$f(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_c) \tag{3.1.7}$$

where Z is a normalization constant. As remarked in section 2.1.2 of [82], the functions ψ_C do not have a clear relationship with the marginal densities of f. The density estimation under a graphical model for general graphs is then too ambitious, and it is necessary to constrain the graph structure.

Forest Density Estimation In a fully nonparametric setting, to our knowledge, one method is available: Forest Density Estimation (FDE) [45]. It corresponds to the estimation of a density with an non-cyclic graphical model (also called a forest).

In this case, the density can be expressed with 1 and 2-dimensional marginals. If $G = (V, E)$ is a forest, the density f of a random variable admitting G as a graphical model enjoys the following formulation

$$f(x) = \prod_{(i,j)\in E} \frac{f_{\{i,j\}}(x_i, x_j)}{f_{\{i\}}(x_i) f_{\{j\}}(x_j)} \prod_{k=1}^{d} f_{\{k\}}(x_k). \tag{3.1.8}$$

In [45] the algorithm to estimate a forest and the corresponding density is presented. Let us emphasis that it requires the estimation of marginals up to dimension 2. Theorem 9 in [45] emphasis that if the true density enjoys a forest graphical model and under suitable condition on the density, the speed of convergence of FDE under Kullback-Leibler (KL) loss is related to the speed of convergence for KDE in dimension 2 instead of in the ambient dimension d. This emphasize that FDE is a remedy to the curse of dimensionality. The KL loss between f and an estimator $\hat{f}$ is defined as

$$\mathrm{KL}\left(f \| \hat{f}\right) = \int \log\left(\frac{f}{\hat{f}}\right) f. \tag{3.1.9}$$

Independence Structure In the present work, we focus on the model of Independence Structure (IS) for multivariate density introduced by Lepski by [42] and also studied by Rebelles [65]. It contains d-dimensional densities, which can be decomposed as a product of low-dimensional marginals, forming a partition of the original features.

$$f(x) = \prod_{S\in\mathcal{P}} f_S(x_S) \tag{3.1.10}$$

Under a graphical model perspective, it corresponds to graphs that are composed of disjoint connected components. Previous works on IS have highlighted that if the density enjoys the property that the size of the biggest block of the partition is equal to $k < d$, then the complexity of density estimation, measured through minimax rate of convergence under L_p losses ($1 \leq p \leq \infty$) is related to k instead of the ambient dimension d. More precisely, it was showed in [65] that

$$\mathrm{R}\left(\|.\|_2^2, \mathcal{H}^\beta(d, H) \cap \mathcal{D}_d^k\right) \asymp N^{-\frac{2\beta}{2\beta+k}}. \tag{3.1.11}$$

However, the estimators introduced by [42] and [65] do not avoid requiring the construction of as many estimators as there are partitions in Part_d^k. This is a

problem if we want to implement their approach for high-dimensional problems, as the cardinality of Part_d^k rapidly become too high for computations.

Kernel Density Estimators In the sequel, we focus on nonparametric density estimation. Kernel Density Estimator (KDE) is a popular density estimator in this context. It has its origins in the works of Rosenblatt [71] and Parzen [59]. It has been successfully used to real-world applications in recent years (connectivity among salmon farms [11], physical activity [36], ecological niche modeling [64], modeling of T cell receptors [57], among many others).

In this chapter, we will consider Spherical Gaussian KDE (SGKDE). For a given bandwidth $h > 0$ we define the SGKDE associated to h and to the sample $X_1, \ldots, X_N$ as

$$\hat{f}_h(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{\exp\left(-\frac{(X_i - x)^{\mathrm{T}}(X_i - x)}{2h^2}\right)}{(2\pi)^{d/2} h^d}. \tag{3.1.12}$$

As we will not consider other choices of kernels, we write KDE instead of SGKDE. The construction of the estimator over a data sample corresponds to the choice of the bandwidth. Different approaches exist. In practice, a cross-validation scheme over a collection of potential values of h is a popular choice. See [41] for analysis in the context of maximum likelihood density estimation.

Although DTM-based density estimators were used in chapter 2 in the density estimation phase of CyToMATo, they are not relevant in the context of this chapter. The issue with g_k^{DTM} is that its integral over $\mathbb{R}^d$ is not equal to one. In that sense it is not a proper density estimator and it has no sense to compare directly g_k^{DTM} with other density estimators under the log-likelihood criterion that we will define later in this chapter.

3.2 ISDE

This section presents ISDE, an algorithm designed to simultaneously perform density estimation and independence partition selection in a moderately high-

dimensional setting.

Specifications Let k be an input parameter. We aim to provide a method taking a point cloud as input and outputting an IS (a partition of the features in Part_d^k) and a density estimator as a product of marginal estimators

$$\hat{f}_{\hat{\mathcal{P}},\hat{h}_{\hat{\mathcal{P}}}} = \prod_{S\in\hat{\mathcal{P}}} \hat{f}_{S,\hat{h}_S} \tag{3.2.1}$$

where $\hat{h}_{\hat{\mathcal{P}}} = \left(\hat{h}_S\right)_{S\in\hat{\mathcal{P}}}$ is a list of bandwidths. For $S \in \text{Set}_d^k$, $\hat{f}_{S,h_S}$ denotes an estimator of the form 3.1.12 constructed with the features belonging to S.

Number of Partitions vs. Number of Subsets Before starting the explanation of how ISDE works, let us highlights some comparison between the number of partitions in Part_d^k and the number of subsets in Set_d^k.

Let us start by comparing S_d and B_d, the respective cardinals of Set_d^d and Part_d^d. We have $S_d = 2^d - 1$ and B_d is known as the Bell number of order d. Table 3.1 shows how these quantities compare for dimension lying between 10 and 15.

d	10	11	12	13	14	15
S_d	$1,023$	$2,047$	$4,095$	$8,191$	$16,383$	$32,767$
B_d	$115,975$	$678,570$	$4,213,597$	$27,644,437$	$190,899,322$	$1,382,958,545$

Table 3.1: Number of partitions vs number of subsets

We remark that the number of partitions is much higher than the number of features. Even if we restrict ourselves to small values of k, the difference remains important. We denote S_d^k and B_d^k the cardinals of Set_d^k and Part_d^k. It is simple to see that

$$S_d^k = \sum_{i=1}^{k} \binom{d}{k}. \tag{3.2.2}$$

For B_d^k exact computation is harder, but a lower bound is computed in lemma 3.2.1.

Lemma 3.2.1

$$B_d^k \geq B_d^2 = 1 + \binom{d}{2} + \frac{\binom{d}{2}\binom{d-2}{2}}{2!} + \frac{\binom{d}{2}\binom{d-2}{2}\binom{d-4}{2}}{3!} \cdots + \frac{\binom{d}{2}\cdots\binom{d-2(\lfloor d/2 \rfloor -1)}{2}}{(\lfloor d/2 \rfloor)!} \tag{3.2.3}$$

Proof. For a nonnegative integer i, let us denote by $B_d^2[i]$ the number of partitions of Part_d^k with exactly i blocks of size 2. A first remark is that $B_d^2[i] = 0$ as soon as $i < \lfloor d/2 \rfloor$, then

$$B_d^2 = \sum_{i=0}^{\lfloor d/2 \rfloor} B_d^2[i]. \tag{3.2.4}$$

Now, we evaluate $B_d^2[i]$. It is not hard to count the number of possibilities to select i pairs of distinct elements of $\{1, \ldots, d\}$ taking into account in which order there were selected. For the first pair, there are $\binom{d}{2}$ choices, then $\binom{d-2}{2}$ choices for selecting another pair among the other variables, and so on. Then there are $\prod_{j=0}^{i-1} \binom{d-2j}{2}$ ordered pairs of variables of $\{1, \ldots, d\}$.

As selecting a partition in Part_d^k is equivalent to an unordered choice of pairs of variables, it remains to divide by the number of permutation of i elements, $i!$. Then

$$B_d^2[i] = \frac{\prod_{j=0}^{i-1} \binom{d-2j}{2}}{i!}. \tag{3.2.5}$$

□

We highlight that $B_d^2 \underset{d\to\infty}{\sim} d^{\frac{d}{2}}$ while $S_d^k \underset{d\to\infty}{\sim} d^k$. For values of d corresponding to moderately high-dimensional settings, some computations are gathered in table 3.2 (the values of B_d^2 are approximations).

d	20	30	40	50
S_d^3	$1,350$	$4,525$	$10,700$	$20,875$
B_d^2	2.4×10^{10}	6.1×10^{17}	7.3×10^{25}	2.8×10^{34}

Table 3.2: Number of partitions vs number of subsets

These computations indicate that it would be beneficial to find a way to avoid the computation of B_d^k estimators. Intuitively, as estimators are combinations of marginals estimators, it seems reasonable to decouple marginal estimations from partition selection. We will now see that we must carefully choose the loss function to implement this idea. As we have indicated in the previous section, the approach proposed by Lepski and Rebelles did not tackle this issue and necessitate to compute S_d^k density estimators, one for each partition, before comparing them.

Choice of Loss Function We have announced in the introduction that ISDE aims to minimize the Kullback-Leibler loss between the proper density and the estimate one. Here we will see that this choice is not innocuous and that other choices of loss function do not lead to a feasible algorithm.

In density estimation literature, the most popular choice for the loss function is undoubtedly the squared L_2 loss. For a partition $\mathcal{P} \in \mathrm{Part}_d^k$ we want to find the collection of bandwidth $(\hat{h}_S^{\mathcal{P}})_{S \in \mathrm{Part}_d^k}$ solutions of

$$\min_{(\hat{h}_S^{\mathcal{P}})_{S \in \mathcal{P}}} \int \left(f - \hat{f}_{\mathcal{P}, h_{\mathcal{P}}} \right)^2 = \int \hat{f}_{\mathcal{P}, h_{\mathcal{P}}}^2 - 2 \int \hat{f}_{\mathcal{P}, h_{\mathcal{P}}} f + \int f^2. \tag{3.2.6}$$

If $P[.]$ corresponds to the integral over the measure induced by the density f, an equivalent formulation is given by

$$\min_{(h_S^{\mathcal{P}})_{S \in \mathcal{P}}} \int \hat{f}_{\mathcal{P}, h_{\mathcal{P}}}^2 - 2P \left[\hat{f}_{\mathcal{P}, h_{\mathcal{P}}} \right] = \min_{(h_S^{\mathcal{P}})_{S \in \mathcal{P}}} \prod_{S \in \mathcal{P}} \int \hat{f}_{S, h_S}^2 - 2P \left[\prod_{S \in \mathcal{P}} \hat{f}_{S, h_S} \right]. \tag{3.2.7}$$

Let $S \in \mathrm{Set}_d^k$ and $\mathcal{P}_1, \mathcal{P}_2 \in \mathrm{Part}_d^k$ such that $S \in \mathcal{P}_1$ and $S \in \mathcal{P}_2$. There is no reason to have $\hat{h}_S^{\mathcal{P}_1} = \hat{h}_S^{\mathcal{P}_2}$ from the previous formulation. Then under the squared L_2 loss, we have no clue on how we can avoid constructing as many estimators as elements in Part_d^k.

Now, for the KL loss, we want to find a collection of bandwidth $(\hat{f}_S)_{S \in \mathrm{Part}_d^k}$ minimizing

$$\min_{(h_S^{\mathcal{P}})_{S \in \mathcal{P}}} \int \log \left(\frac{f}{\hat{f}_{\mathcal{P}, h_{\mathcal{P}}}} \right) f. \tag{3.2.8}$$

An equivalent formulation is given by

$$\max_{(h_S^{\mathcal{P}})_{S\in\mathcal{P}}} P\left[\log \hat{f}_{\mathcal{P},h_{\mathcal{P}}}\right] = \max_{(h_S^{\mathcal{P}})_{S\in\mathcal{P}}} \sum_{S\in\mathcal{P}} \left\{P\left[\log \hat{f}_{S,h_S}\right]\right\} \tag{3.2.9}$$

using the property that the logarithm changes products into sums and the linearity of the operator $P[.]$. By opposition of what we have seen for the squared L_2 loss, if $S \in \mathcal{P}_1$ and $S \in \mathcal{P}_2$, we will have $h_S^{\mathcal{P}_1} = h_S^{\mathcal{P}_2}$. Then under KL loss, bandwidths optimization over marginal estimators and partition selection can be decoupled, leading to the necessity of computing S_d^k density estimators instead of B_d^k. As shown in table 3.1 and table 3.2, it leads to an appreciable gain in terms of algorithmic complexity.

Empirical Formulation of the Optimization Problem Under KL loss, bandwidths optimization and partition selection become two separated tasks. This decoupling incites us to design an algorithm consisting of two steps: first, compute a marginal estimator $\hat{f}_S$ for all $S \in \mathrm{Set}_d^k$ and then find the best combination of them for a log-likelihood criterion. Let n and m be two positive integers, such that $m + n = N$. The dataset $X_1, \ldots, X_N$ is split into two disjoint sub-samples:

- $W_1, \ldots, W_m$ used to compute marginal estimators $(\hat{f}_S)_{S\in\mathrm{Set}_d^k}$
- $Z_1, \ldots, Z_n$ used to compute empirical log-likelihoods $(\ell_n(S))_{S\in\mathrm{Set}_d^k}$ where $\ell_n(S) = \frac{1}{n}\sum_{i=1}^n \log\left(\hat{f}_S(Z_i)\right)$

Let us use the notation $\ell_n(\mathcal{P}) = \sum_{S\in\mathcal{P}} \ell_n(S)$. The empirical optimization task can be written as

$$\max_{\mathcal{P}\in\mathrm{Part}_d^k} \ell_n(\mathcal{P}) = \max_{\mathcal{P}\in\mathrm{Part}_d^k} \sum_{S\in\mathcal{P}} \ell_n(S). \tag{3.2.10}$$

Partition Selection A naive approach to solve 3.2.10 is to compute $\ell_n(\mathcal{P})$ for every partition of Part_d^k and then find the optimal one. However, this approach becomes time-consuming when d grows and infeasible for large values of d because

of the number of partitions. Therefore, it will be appreciable to reformulate this optimization to speed up computation. It is possible to reformulate 3.2.10 as the following linear programming task.

Solve

$$\max_{x \in \mathbb{R}^{\mathrm{Set}_d^k}} \sum_{S \in \mathrm{Set}_d^k} \ell_n(S) x(S) \tag{3.2.11}$$

Under constraints

$$Ax = (1, \ldots, 1)^{\mathrm{T}} \tag{3.2.12}$$

$$x \in \{0,1\}^{S_d^k}. \tag{3.2.13}$$

Where x is a binary vector representing which elements of Set_d^k are selected, and A is a $d \times S_d^k$ matrix where each column is a binary vector representing the composition of one of the sets of Set_d^k. The condition $Ax = (1, \ldots, 1)^{\mathrm{T}}$ then ensures that each feature is chosen once, implying that the sets selected with x form a partition.

We validate this approach through a running time comparison (see table 3.3) between the implementation of a brute-force approach and a linear program solver. In this experiment, we fix the quantities $(\ell_n(S))_{S \in \mathrm{Set}_d^k}$, the brute-force approach consists in a for loop (implemented in Python), computing $\ell_n(\mathcal{P})$ for all $\mathcal{P} \in \mathrm{Part}_d^k$ and returning the maximum. For the LP formulation, the optimization is done with the branch-and-bound method, implemented in the Python package PulP [52]. With the brute-force approach and choice $k = d$, partition selection takes approximately 3 hours in dimension 15 but less than 10 seconds with LP formulation.

d	9	10	11	12	13	14	15
Brute-Force Approach	0.2	0.9	5.2	32	219	1304	10437
LP Solver	0.1	0.2	0.4	0.8	1.9	4.1	9.1

Table 3.3: Running time (seconds): linear programming vs brute-force approach for partition selection

Conclusion The resulting algorithm is algorithm 2. It enjoys the following properties:

- It exploits the decoupling of marginal density estimation and partition selection offered by choice of KL as discrepancy measure: it optimizes over partitions in Part_d^k even if it only requires the computation of Set_d^k marginal estimators.
- It is versatile: even if we present the construction of ISDE using KDEs for marginal estimation, it is possible to use any other base multivariate density estimator.

input : $X_1, \ldots, X_N \in \mathbb{R}^d$, k integer with $k \leq d$, integers m and n and a subroutine to perform multidimensional density estimation
output: Partition $\hat{\mathcal{P}} \in \text{Part}_d^k$, marginal estimates $(\hat{f}_S)_{S \in \hat{\mathcal{P}}}$
begin
  for $S \in \text{Set}_d^k$ **do**
    Compute $\hat{f}_S(W_1, \ldots, W_m)$ thanks to the density estimation subroutine
    Compute $\ell_n(S)$
  end
  Compute $\hat{\mathcal{P}} \in \underset{\mathcal{P} \in \text{Part}_d^k}{\arg\max} \sum_{S \in \mathcal{P}} \ell_n(S)$ using linear programming formulation
end

Algorithm 2: ISDE

3.3 Experiments on synthetic data

In this section, we validate the performance of ISDE on synthetic data generated under IS hypothesis.

Data Generating Process For a given list of positive integer (a structure) $S = [s_1, \ldots, s_K]$, the data generating process is defined as follows. For each $s_i \in S$, we define a s_i dimensional dataset drawn from P_i:

- If $s_i = 1$, P_i is the uniform distribution over $[0, 1]$

- If $s_i = 2$, P_i is a distribution corresponding to data sample near two concentric circles with different radii
- If $s_i = 3$, a sample X from P_i is obtained as follows: let Y_1 and Y_2 be two independent Bernoulli variables with probability of success 0.5 and $Y_3 = |Y_1 - Y_2|$. X is then drawn from the multivariate Gaussian distribution $\mathcal{N}((Y_1, Y_2, Y_3), 0.08 \times I_3)$. This is a situation where features of P_i are pairwise independent but not mutually independent
- If $s_i \geq 4$, P_i is a mixture of two multivariate Gaussian distributions, one centered in $(0, \ldots, 0)$, the other in $(1, \ldots, 1)$

The final dataset results from their concatenation, plus feature-wise rescaling so that each value lies between 0 and 1. The dimension is $d = \sum_{i=1}^{k} s_i$. This rescaling step does not affect the IS as it is done feature-wise.

Evaluation Scheme To evaluate the performance of an estimator, we compute the empirical log-likelihood on a validation set $X^{\text{valid}} = X_1^{\text{valid}}, \ldots, X_M^{\text{valid}}$ drawn independently of the same distribution as $X_1, \ldots, X_N$:

$$\text{Score}(\hat{f}) = \frac{1}{M} \sum_{i=1}^{M} \log\left(\hat{f}\left(X_i^{\text{valid}}\right)\right). \tag{3.3.1}$$

The set $X^{\text{valid}} = X_1^{\text{valid}}, \ldots, X_M^{\text{valid}}$ is not used to tune the estimators. In the experiments of this section, we set $M = 5000$.

Benchmarked Methods We will compare three density estimation algorithms for samples corresponding to different structures.

The first one is CVDKE, a KDE estimator where the bandwidth parameter is selected through a 5-fold cross-validation to maximize empirical log-likelihood on test data. The collection of possible bandwidths is a regular grid on a log-scale from 0.01 to 1 with 30 values.

The second one is ISDE with $k = d$ (*ie* all partitions are tested), $m = n = 0.5N$ and the collection of marginal estimators $(\hat{f}_S)_{S\in \mathrm{Set}_d}$ is a collection of CVKDE estimators constructed with the sample $W_1, \ldots, W_m$.

The third one is FDE. Our implementation is a slight modification of the held-out data approach proposed in [45]: we rely on the quantities $(\ell_n(S))_{S\in \mathrm{Set}_d^2}$ computed in ISDE as estimators of the quantities $\left(\int \log(f_S) f_S\right)_{S\in \mathrm{Set}_d^2}$. We use a cross-validation scheme to optimize the bandwidth instead of the plug-in approach presented in the paper.

We insist that comparing these methods for density estimation through empirical log-likelihood for validation data is fair, as all of them aim to maximize the log-likelihood.

Results Empirical log-likelihood on validation data for methods listed above are shown in table 3.4, for different structures and for the choice $N = 5000$. Each experiment is repeated 5 times, and we show the mean log-likelihood and the standard deviation on the table.

	[2, 2, 1]	[3, 3, 3]	[4, 4, 2, 2]
ISDE	$\mathbf{1.83 \pm 0.08}$	$\mathbf{4.05 \pm 0.15}$	$\mathbf{6.30 \pm 0.25}$
FDE	$\mathbf{1.83 \pm 0.08}$	2.88 ± 0.14	5.89 ± 0.33
CVKDE	0.56 ± 0.03	3.49 ± 0.11	3.96 ± 0.16

Table 3.4: Empirical log-likelihood on validation data for different density estimators

Conclusion For $[2, 2, 1]$, ISDE and FDE give similar results as they output the same graph and the same bandwidths. They both outperform CVKDE. For $[3, 3, 3]$, as features are pairwise independent, FDE outputs at every try a graph without any edge and computes the density as a product of one-dimensional marginals, leading to poor results in comparison to ISDE. CVKDE leads to better estimation for this setting than FDE, but is outperformed by ISDE. For $[4, 4, 2, 2]$, FDE outputs a sub-graph of the actual graphical model at every try. It leads to better estimation than CVKDE but worse than ISDE, which learns the proper IS at every try.

Thus, ISDE leads to better results than FDE and CVKDE for the task of structured density estimation under KL loss under IS. We interpret the bad performance of CVKDE as a manifestation of the curse of dimensionality. ISDE outperforms FDE because it considers potential higher-order dependencies between features than FDE, which only considers pairwise associations. However, let us remark that FDE covers some models not addressed by ISDE. ISDE performs better on data where IS is true, but we recommend testing both methods to determine the one that best fits the data.

We also remark that ISDE recovers exactly the IS for the considered settings. One can wonder why we do not observe that outputted partitions are not precisely the IS, but partitions where blocks are a union of blocks of the proper IS. We believe that this is because a useless merging of blocks in the partition is strongly penalized by ISDE as the dimension limits our ability to estimate a density accurately. Then the hold-out scheme implemented in ISDE (by splitting X into W and Z in algorithm 2) penalizes sufficiently too big blocks in partitions and leads to accurate recovery of IS.

3.4 Complexity and running time analysis

In this section, we provide information about the algorithmic complexity and running time of ISDE.

Computation of KDE For a given bandwidth h, the evaluation of a KDE constructed over m_1 points and evaluated over m_2 points is $O(m_1 m_2)$. The family of estimators $\left(\hat{f}_S\right)_{S\in \mathrm{Set}_d^k}$ is constructed using a V-fold cross-validation where V is a divisor of m. If n_h denotes the number of candidate values for the bandwidths, the number of operation required for bandwidth selection is $S_d^k n_h V \frac{m}{V} \times \frac{m(V-1)}{V}$. The complexity of this step is $O(S_d^k n_h m^2)$. Once the bandwidths are selected, it remains to compute the quantities $(\ell_n(S))_{S\in \mathrm{Set}_d^k}$ thanks to $Z_1, \ldots, Z_n$. The total cost of its operation is $O(S_d^k nm)$. The total algorithmic cost of the computation of $(\ell_n(S))_{S\in \mathrm{Set}_d^k}$ is

$$O\left(S_d^k m\left(n_h m + n\right)\right). \tag{3.4.1}$$

Partition Selection The implementation of the partition selection step relies on the branch-and-bound method. It is not easy to give a precise statement about its complexity. The branch-and-bound algorithm uses a tree search strategy to enumerate all possible solutions to a given problem implicitly. A recent survey can be found in [54].

Running time We now present some information about running time. We have run all experiments on a laptop with the following hardware: CPU Intel Xeon W-10885M CPU @ 2.40GHz and GPU: Nvidia Quadro RTX 3000 Mobile.

The KDE computations have been performed on GPU using the python package KeOps [12]. This implementation is much faster than the one on CPU proposed by scikit-learn [60] as highlighted by table 3.5, which compares running time for KDE constructed on n points and evaluated on n points on dimension $d = 3$.

n	100	500	2000	$5,000$	$10,000$	$20,000$
KeOps implementation	0.0006	0.0020	0.0073	0.0176	0.0684	0.1163
Scikit-learn implementation	0.0008	0.0126	0.1952	1.1564	4.9151	21.3631

Table 3.5: Comparison of running time (seconds) of sklearn implementation and ours for KDE constructed on n points and evaluated on n points

The computation of the quantities $(\ell_n(S))_{S \in \mathrm{Set}_d^k}$ requires many repetitions of KDE evaluation. In table 3.6 we provide estimation of the running time for this step for various values of k and d and considering a 5-fold cross-validation to estimate each bandwidth among 30 candidate values. The quantities m and n are both set to $1,000$.

k \ d	5	10	20	30	40	50
2	2.0	6.8	18	40	69.15	108
3	3.3	23	121	409	949	1,862
4	3.9	53	590	3,053	9,572	23,536
5	4.2	61	2,154	17,589	75,228	233,323

Table 3.6: Running time (seconds) for $(\ell_n(S))_{S\in \mathrm{Set}_d^k}$ computation with respect to k and d and with 5-fold cross selected bandwidths over 30 possible values and for $m = n = 1000$

Once the quantities $(\ell_n(S))_{S\in \mathrm{Set}_d^k}$ are computed, it remains to perform partition selection. As mentioned previously, we use the python package Pulp [52]. The running time of this step for different values of k and d are presented in table 3.7.

k \ d	5	10	20	30	40	50
2	0.02	0.03	0.08	0.20	0.47	0.84
3	0.02	0.05	0.41	1.9	6.2	15
4	0.02	0.09	2.0	14.8	62	190
5	0.02	0.13	7.3	84.7	482	2,045

Table 3.7: Running time (seconds) for partition selection step with respect to k and d

The main conclusion of this running time study is that the running time of partition selection is negligible in comparison with the one for computing $(\ell_n(S))_{S\in \mathrm{Set}_d^k}$ for the parameters presented here. The code associated with this chapter contains functions allowing the reader to reproduce these experiments with different settings and estimate the running time on its device. Note that the code also runs if no GPU is available. In this case, KeOps will automatically use parallelization on CPU for KDE evaluations.

3.5 Conclusion

ISDE is an algorithm that outputs an estimate of a density function of a point cloud, taking into account an IS for data in moderately high dimensions. To design it, we reduced the number of hyper-parameters with an appropriate choice of the loss function and, through linear programming reformulation, made the partition selection step faster than was previously possible. This leads to reasonable running time even on a laptop for the considered datasets. The code is available and ready to be used by anyone interested in this method.

ISDE is versatile: it takes any basic multidimensional density estimator as input. Then it can be used in parametric and nonparametric frameworks. It is also exhaustive as it searches over all partitions of features with given maximal block size. To our knowledge, we are the first to propose a scalable algorithm that considers IS in the context of nonparametric density estimation with KDE for moderately high-dimensional data.

We validated its performance on synthetic data satisfying IS. This performance was measured in terms of log-likelihood on the validation sample. We found that ISDE exploits IS structure and outperform other density estimators for this task. Note that we restricted ourselves in this chapter to the nonparametric case, as it is the version of ISDE that we will use for analyzing cytometry data. A study similar to the one presented in section 3.3 but for multivariate Gaussian data can be found in appendix A.

Code availability The code to reproduce the experiments presented here is available at https://github.com/Louis-Pujol/ISDE-Paper.

Chapter 4

An upper-bound of the Kullback-Leibler risk of the density estimated by ISDE

Context Obtaining upper-bounding inequality for learning algorithm is a central question in statistical learning. In the context of density estimation, a vast literature have dealt with the question of finding an upper-bound for the expectation of the L_p loss. These bounds are instructive as they indicate how the quality of estimation is related with the hypothesis on the true density. When available, lower bounds permit to fully characterized the estimation problem. In the context of Kullback-Leibler loss, there is significantly less available literature. The reason is because KL loss is analytically less convenient than L_p losses, especially when densities can take values close to zero. With the additional hypothesis that all densities are bounded away from zero, KL loss is equivalent to other classical losses for density estimation such as Hellinger, Kullback-Jensen and squared L_2. However, in the context of ISDE, we cannot just add a boundary hypothesis on the unknown density and switch to a more convenient loss, because the combinatorial complexity reduction operated by ISDE is intimately related to the use of the KL loss.

This chapter is intended to provide a theoretical analysis of ISDE by upper-bounding the quantity $\mathrm{KL}\left(f \| \hat{f}_{\hat{\mathcal{P}}}\right)$ where $\hat{f}_{\hat{\mathcal{P}}}$ is the output of ISDE. In particular,

we will show that the introduction of IS tackles the curse of dimensionality and that the constants in the upper-bound reflect the combinatorial complexity reduction implemented in ISDE.

Our contribution In this chapter, we show an upper-bound for the Kullback-Leibler loss of the estimator outputted by ISDE and the true density valid under some conditions. We impose an β-Hölder regularity condition on the marginal of the unknown density, as well as a condition controlling that the marginal densities are greater than a positive quantity. Then we show that for a well-chosen family of marginal density estimators, namely mirror-images KDE, if the true density actually enjoys an Independence Structure with blocks of size at most k, then with probability $(1-2/m)(1-2/n)$(see theorem 4.4.1)

$$\begin{aligned}\mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) \leq & C_1\sqrt{\log m + \log\left(S_d^k\right)}\left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+k}} \\ & + C_2\sqrt{\log n + \log\left(S_d^k\right)}\frac{k}{\sqrt{n}} \end{aligned} \tag{4.0.1}$$

where C_1 and C_2 are constants independent of the sample sizes m and n. This bound shows that ISDE tackles the curse of dimensionality as the speed of convergence is related to k instead of the ambient dimension d and highlights the combinatorial complexity reduction operated by ISDE as S_d^k, the number of subsets of variables of size k appears as a complexity parameter.

Organization of the chapter In section 4.1 we establish a first decomposition on the risk involving oracle partitions. In section 4.2 we introduce the regularity conditions on the proper density and establish that an upper-bound for the uniform loss between marginal densities of f and marginal estimators is sufficient to obtain a convergence result for ISDE. In section 4.3 we show that it is possible to obtain an upper-bound for uniform estimation of marginal densities for a particular estimator. Then in section 4.4 we establish the desired upper-bound for the estimator outputted by ISDE.

4.1 Kullback-Leibler risk decomposition

In this section, we show that the KL loss between f and $\hat{f}_{\hat{\mathcal{P}}}$, the estimator outputted by ISDE, decomposes as the sum of three terms with a clear interpretation.

4.1.1 Oracles partitions

We denote by $\hat{\mathcal{P}}$ the partition outputted by ISDE. Let $P_n(.)$ denotes the empirical measure associated with the sample $Z_1, \ldots, Z_n$ and $P(.)$ the measure associated with the true density f. For any measurable function g we have

$$P(g) = \int g(x)f(x)dx \tag{4.1.1}$$

and

$$P_n(g) = \frac{1}{n}\sum_{i=1}^{N} g(Z_i). \tag{4.1.2}$$

$\hat{\mathcal{P}}$ is solution of the following optimization problem :

$$\hat{\mathcal{P}} \in \underset{\mathcal{P}\in \mathrm{Part}_d^k}{\arg\min}\ P_n\left(-\log(\hat{f}_{\mathcal{P}})\right) \tag{4.1.3}$$

The partition $\hat{\mathcal{P}}$ is random, depending on both W and Z. Let us define two other meaningful partitions.

$$\tilde{\mathcal{P}} \in \underset{\mathcal{P}\in \mathrm{Part}_d^k}{\arg\min}\ P\left(-\log(\hat{f}_{\mathcal{P}})\right) = \underset{\mathcal{P}\in \mathrm{Part}_d^k}{\arg\min}\ \mathrm{KL}\left(f\|\hat{f}_{\mathcal{P}}\right) \tag{4.1.4}$$

and

$$\mathcal{P}_* \in \underset{\mathcal{P}\in \mathrm{Part}_d^k}{\arg\min}\ P\left(-\log(f_{\mathcal{P}})\right) = \underset{\mathcal{P}\in \mathrm{Part}_d^k}{\arg\min}\ \mathrm{KL}\left(f\|f_{\mathcal{P}}\right). \tag{4.1.5}$$

$\tilde{\mathcal{P}}$ is a random partition depending on W but not on Z. It is the best combination of the estimators $(\hat{f}_S)_{S\in \mathrm{Set}_d^k}$ if we consider that the quantities $\left(P(-\log \hat{f}_S)\right)_{S\in \mathrm{Set}_d^k}$ are known. $\mathcal{P}_*$ is not random. It is only a function of k. $f_{\mathcal{P}_*}$ can be interpreted as the Kullback-Leibler projection of f on the model $\mathcal{D}_d^k$ thanks to the following property.

Proposition 4.1.1

$$f_{\mathcal{P}_*} \in \arg\min_{g\in\mathcal{D}_d^k} \mathrm{KL}\left(f\|g\right) \tag{4.1.6}$$

Proof. Let $g \in \mathcal{D}_d^k$ and denote by $\mathcal{P}_g$ a partition such that $g = \prod_{S\in\mathcal{P}_g} g_S$. We have

$$\begin{aligned}
\mathrm{KL}\left(f\|g\right) &= \int \log\left(\frac{f}{g}\right) f && (4.1.7)\\
&= \int \log\left(\frac{f}{f_{\mathcal{P}_g}}\right) f + \int \log\left(\frac{f_{\mathcal{P}_g}}{g}\right) f && (4.1.8)\\
&= \mathrm{KL}\left(f\|f_{\mathcal{P}_g}\right) + \sum_{S\in\mathcal{P}_g} \mathrm{KL}\left(f_S\|g_S\right) && (4.1.9)\\
&\geq \mathrm{KL}\left(f\|f_{\mathcal{P}_g}\right) && (4.1.10)
\end{aligned}$$

with equality if $g = f_{\mathcal{P}_g}$. Then

$$\begin{aligned}
\min_{g\in\mathcal{D}_d^k} \mathrm{KL}\left(f\|g\right) &= \min_{\mathcal{P}\in\mathrm{Part}_d^k} \left(\min_{g\in\mathcal{D}_d^k} \mathrm{KL}\left(f\|g\right)\right) && (4.1.11)\\
&= \min_{\mathcal{P}\in\mathrm{Part}_d^k} \mathrm{KL}\left(f\|f_{\mathcal{P}}\right) && (4.1.12)
\end{aligned}$$

□

4.1.2 Kullback-Leibler risk upper-bound

We are now in a position to establish a control of the Kullback-Leibler risk for $\hat{f}_{\hat{P}}$ involving the oracles partitions.

Lemma 4.1.2

$$\mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) \leq \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right) + \sum_{S_*\in\mathcal{P}_*} \mathrm{KL}\left(f_{S_*}\|\hat{f}_{S_*}\right) + (P-P_n)(\log \hat{f}_{\tilde{\mathcal{P}}} - \log \hat{f}_{\hat{\mathcal{P}}}) \tag{4.1.13}$$

Proof. We start by decomposing $\mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right)$ as follows

$$\begin{aligned}
\mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) &= \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right) && (4.1.14)\\
&+ \mathrm{KL}\left(f\|\hat{f}_{\mathcal{P}_*}\right) - \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right) && (4.1.15)\\
&+ \mathrm{KL}\left(f\|\hat{f}_{\tilde{\mathcal{P}}}\right) - \mathrm{KL}\left(f\|\hat{f}_{\mathcal{P}_*}\right) && (4.1.16)\\
&+ \mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) - \mathrm{KL}\left(f\|\hat{f}_{\tilde{\mathcal{P}}}\right). && (4.1.17)
\end{aligned}$$

Then, as $\mathrm{KL}\left(f\|\hat{f}_{\tilde{\mathcal{P}}}\right) \leq \mathrm{KL}\left(f\|\hat{f}_{\mathcal{P}_*}\right)$,

$$\begin{aligned}
\mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) &\leq \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right) && (4.1.18)\\
&+ \mathrm{KL}\left(f\|\hat{f}_{\mathcal{P}_*}\right) - \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right) \qquad \text{(i)} && (4.1.19)\\
&+ \mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) - \mathrm{KL}\left(f\|\hat{f}_{\tilde{\mathcal{P}}}\right) \qquad \text{(ii)}. && (4.1.20)
\end{aligned}$$

Now, we rewrite (i)

$$\begin{aligned}
\mathrm{KL}\left(f\|\hat{f}_{\mathcal{P}_*}\right) - \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right) &= \int \log\left(\frac{f(x)}{\hat{f}_{\mathcal{P}_*}(x)}\right) f(x)dx - \int \log\left(\frac{f(x)}{f_{\mathcal{P}_*}(x)}\right) f(x)dx && (4.1.21)\\
&= \int \log\left(\frac{f_{\mathcal{P}_*}(x)}{\hat{f}_{\mathcal{P}_*}(x)}\right) f(x)dx && (4.1.22)\\
&= \sum_{S_*\in\mathcal{P}_*} \int \log\left(\frac{f_{S_*}(x)}{\hat{f}_{S_*}(x)}\right) f_{S_*}(x)dx && (4.1.23)\\
&= \sum_{S_*\in\mathcal{P}_*} \mathrm{KL}\left(f_{S_*}\|\hat{f}_{S_*}\right). && (4.1.24)
\end{aligned}$$

And we upper-bound (ii). As $P_n\left[\log \hat{f}_{\hat{\mathcal{P}}} - \log \hat{f}_{\tilde{\mathcal{P}}}\right] \geq 0$:

$$\begin{aligned}
\mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) - \mathrm{KL}\left(f\|\hat{f}_{\tilde{\mathcal{P}}}\right) &= P\left[-\log \hat{f}_{\hat{\mathcal{P}}}\right] - P\left[-\log \hat{f}_{\tilde{\mathcal{P}}}\right] && (4.1.25)\\
&\leq P\left[\log \hat{f}_{\tilde{\mathcal{P}}} - \log \hat{f}_{\hat{\mathcal{P}}}\right] + P_n\left[\log \hat{f}_{\hat{\mathcal{P}}} - \log \hat{f}_{\tilde{\mathcal{P}}}\right] && (4.1.26)\\
&= (P - P_n)(\log \hat{f}_{\tilde{\mathcal{P}}} - \log \hat{f}_{\hat{\mathcal{P}}}). && (4.1.27)
\end{aligned}$$

□

Three terms appear in the upper bound, and they can be easily interpreted.

- $\mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right)$ is a bias term. It is the intrinsic error of the model $\mathcal{D}_d^k$ and can be thought of as a distance from f to $\mathcal{D}_d^k$ thanks to proposition 4.1.1.
- $\sum_{S_*\in\mathcal{P}_*} \mathrm{KL}\left(f_{S_*}\|\hat{f}_{S_*}\right)$ is an approximation term. It is a random quantity depending on the sample W and represents the error made when $f_{\mathcal{P}_*}$ is estimated with $\hat{f}_{\mathcal{P}_*}$.
- $(P - P_n)(\log \hat{f}_{\tilde{\mathcal{P}}} - \log \hat{f}_{\hat{\mathcal{P}}})$ is a selection term. It depends on both W and Z. Conditionally to Z, it quantifies the discrepancy between the integral of the quantities $(\hat{f}_{\mathcal{P}})_{\mathcal{P}\in \mathrm{Part}_d^k}$ and their estimation with P_n. Conditionally to W, it depends on how the estimation of log-likelihoods made thanks to Z is accurate and quantifies our ability to output the optimal partition.

In the sequel of the chapter, we will focus on upper-bounding the approximation and selection terms, as they are the random quantities of interest in our problem. We treat the bias term as a structural error, and we focus on upper-bounding the quantity

$$\mathrm{KL}\left(f\|\hat{f}_{\hat{\mathcal{P}}}\right) - \mathrm{KL}\left(f\|f_{\mathcal{P}_*}\right). \tag{4.1.28}$$

Obtaining such a bound requires to impose some restrictions on the true density f. Here we will consider that f is compactely supported and lower-bounded by a positive value. In this setting, we were not able to obtain precise statements for the bias term. However, in the multivariate Gaussian framework, a precise study of the bias is possible. As this model falls outside the scope of the present chapter, we postpone the study of the bias in a multivariate Gaussian framework to appendix A.

4.2 Conditions on the true density and objective

4.2.1 Regularity conditions

Bounding condition Density estimation under Kullback-Leibler loss is known to be a challenging problem. One work by [31] has studied the asymptotic convergence rates for kernel estimators in a one-dimensional setting. It was shown that the tails of the kernel must be chosen appropriately regarding the tails property of the proper density to have convergent estimators. In this work, we restrict our attention to densities that are lower and upper bounded by some positive quantities. This is done to avoid hardly tractable tail behavior issues. In the sequel, we consider that the following bounding condition is valid for all $S \in \mathrm{Set}_d^k$

$$e^{-A|S|} \leq f_S \leq e^{A|S|} \ \forall S \in \mathcal{P}_* \tag{BC}$$

This condition is a way to ensure that f_S is lower and upper-bounded by some positive quantities. The specific form of the bounds we have chosen in (BC) is a way to simplify the results in the remainder of the chapter, as we will deal with log-densities. Note that if we impose a positive lower bound on the marginal densities, we must consider that f is compactly supported. In the sequel, we will suppose that the support of f is $[0,1]^d$.

Hölder Regularity We will consider in the sequel that it exists $\beta \in (0,2]$ and $L > 0$ such that $f_S \in \mathcal{H}(\beta, L)$ for all $S \in \mathrm{Set}_d^k$. We will use the following approximation property for functions in Hölder balls.

Lemma 4.2.1: Lemma 1 in [87]

Let us consider that $g \in \mathcal{H}(\beta, L)$ with $\beta \in (0, 2)$ and the domain of g is $\mathcal{U} \subset \mathbb{R}^d$, then for all $x \in \mathcal{U}$ and u such that $x + u \in \mathcal{U}$. If $\beta \in (0, 1]$ then

$$|g(x) - g(x + u)| \leq L\|u\|^\beta. \tag{4.2.1}$$

If $\beta \in (1, 2]$ then

$$\left| g(x) - g(x + u) - \sum_{k=1}^{d} \partial_k u_k g(x) \right| \leq L\|u\|^\beta. \tag{4.2.2}$$

4.2.2 Objective

Our goal is to propose an estimation procedure for the collection of marginal densities $(f_S)_{S \in \mathrm{Set}_d^k}$. If we are able to ensure, simultaneously for all $S \in \mathrm{Set}_d^k$ a uniform control of the form

$$\|\hat{f}_S - f_S\|_\infty \leq \epsilon_S < e^{-A|S|}(1 - e^{-A|S|}) \tag{UC}$$

then we can upper-bound the approximation term and the selection term thanks to the following proposition.

Proposition 4.2.2

If the uniform control (UC) is satisfied and (BC) is true, then

1. A bounding condition is satisfied by all the estimators $(\hat{f}_S)_{S\in \mathrm{Set}_d^k}$

$$e^{-2A|S|} \leq \hat{f}_S \leq e^{2A|S|}. \tag{$\widehat{\mathrm{BC}}$}$$

2. For all $S \in \mathrm{Set}_d^k$, the Kullback-Leibler divergence between f_S and $\hat{f}_S$ can be upper-bounded

$$\mathrm{KL}\left(f_S \| \hat{f}_S\right) \leq e^{2A|S|}\epsilon_S \tag{4.2.3}$$

3. Conditionally to W, the selection term can be upper-bounded with high probability. More precisely, if $\delta \in (0,1)$ we have

$$\mathbb{P}\left[\left|(P-P_n)(\log \hat{f}_{\tilde{\mathcal{P}}} - \log \hat{f}_{\hat{\mathcal{P}}})\right| \geq 4\sqrt{2}d\frac{Ak}{\sqrt{n}}\sqrt{\log\left(\frac{2S_d^k}{\delta}\right)}\,\middle|\, W\right] \leq \delta. \tag{4.2.4}$$

Proof. Proof of 1: Under (BC) we have

$$e^{-A|S|} - \|\hat{f}_S - f_S\|_\infty \leq \hat{f}_S \leq e^{A|S|} + \|\hat{f}_S - f_S\|_\infty. \tag{4.2.5}$$

Now, under (UC)

$$e^{-A|S|} - \|\hat{f}_S - f_S\|_\infty \geq e^{-A|S|} - e^{-A|S|}(1-e^{-A|S|}) = e^{-2A|S|} \tag{4.2.6}$$

and

$$\begin{aligned} e^{A|S|} + \|\hat{f}_S - f_S\|_\infty &\leq e^{A|S|} + e^{-A|S|}(1-e^{-A|S|}) && (4.2.7)\\ &\leq e^{A|S|} + e^{3A|S|}\left(e^{-A|S|}(1-e^{-A|S|})\right) && (4.2.8)\\ &= e^{A|S|} + e^{2A|S|}(1-e^{-A|S|}) = e^{2A|S|}. && (4.2.9) \end{aligned}$$

Proof of 2: Let us compute

$$\begin{aligned} \mathrm{KL}\left(f_S \| \hat{f}_S\right) &= \int \log\left(\frac{f_S}{\hat{f}_S}\right) f_S && (4.2.10)\\ &\leq \int \left(\frac{f_S - \hat{f}_S}{\hat{f}_S}\right) f_S && (4.2.11) \end{aligned}$$

Using $(\widehat{\mathrm{BC}})$, $1 \,/\, \hat{f}_S \leq e^{2A|S|}$

$$\leq e^{2A|S|}\|f_S - \hat{f}_{S,h_m}\|_\infty \tag{4.2.12}$$

$$\leq e^{2A|S|}\epsilon_S \tag{4.2.13}$$

Proof of 3: Let $S \in \mathrm{Set}_d^k$, under $(\widehat{\mathrm{BC}})$ we have $\log \hat{f}_S \in [-2A|S|, 2A|S|]$. Using Hoeffding inequality (theorem 2.8 of [8]), we obtain

$$\mathbb{P}\left[\left|(P - P_n)\log \hat{f}_S\right| \geq \frac{2\sqrt{2}A|S|}{\sqrt{n}}\sqrt{\log \frac{2S_d^k}{\delta}}|W\right] \leq \frac{\delta}{S_d^k}. \tag{4.2.14}$$

Now, by union bound :

$$\mathbb{P}\left[\sup_{S\in\mathrm{Set}_d^k}\left|(P - P_n)\log \hat{f}_S\right| \geq \frac{2\sqrt{2}A|S|}{\sqrt{n}}\sqrt{\log \frac{2S_d^k}{\delta}}|W\right] \leq \delta. \tag{4.2.15}$$

This leads to :

$$\mathbb{P}\left[2d\sup_{S\in\mathrm{Set}_d^k}\left|(P - P_n)\log \hat{f}_S\right| \leq 2d\frac{2\sqrt{2}A|S|}{\sqrt{n}}\sqrt{\log \frac{2S_d^k}{\delta}}|W\right] \geq 1 - \delta. \tag{4.2.16}$$

Now, we remark that

$$|(P - P_n)(\log \hat{f}_{\tilde{\mathcal{P}}} - \log \hat{f}_{\hat{\mathcal{P}}})| = \left|\sum_{S\in\tilde{\mathcal{P}}}(P - P_n)\log \hat{f}_S - \sum_{S\in\hat{\mathcal{P}}}(P - P_n)\log \hat{f}_S\right| \tag{4.2.17}$$

$$\leq \sum_{S\in\tilde{\mathcal{P}}}\left|(P - P_n)\log \hat{f}_S\right| + \sum_{S\in\hat{\mathcal{P}}}\left|(P - P_n)\log \hat{f}_S\right| \tag{4.2.18}$$

$$\leq 2d\sup_{S\in\mathrm{Set}_d^k}\left|(P - P_n)\log \hat{f}_S\right|. \tag{4.2.19}$$

Then, we have

$$\mathbb{P}\left[\left|(P - P_n)(\log \hat{f}_{\tilde{\mathcal{P}}} - \log \hat{f}_{\hat{\mathcal{P}}})\right| \geq 4\sqrt{2}d\frac{Ak}{\sqrt{n}}\sqrt{\log\left(\frac{2S_d^k}{\delta}\right)}\,\middle|\, W\right] \leq \delta. \tag{4.2.20}$$

□

4.3 Uniform density estimation for marginal densities

4.3.1 For a fixed S

In this subsection, we fix a subset of variables $S \in \mathrm{Set}_d^k$, and we study the problem of constructing an estimator $\hat{f}_S$ based on the sample $W_1, \dots, W_m$ giving a control of $\|f_S - \hat{f}_S\|_\infty$ in order to verify (UC). We decompose the error as a sum of a bias and a variance term as follows

$$\|f_S - \hat{f}_S\|_\infty \leq \underbrace{\left\|f_S - \mathbb{E}\left[\hat{f}_S\right]\right\|_\infty}_{\text{Bias}} + \underbrace{\left\|\mathbb{E}\left[\hat{f}_S\right] - \hat{f}_S\right\|_\infty}_{\text{Variance}}. \tag{4.3.1}$$

Bias upper-bound

Choice of the kernel function In the following, we will use a density estimator based on an ancillary function K called kernel. K is a nonnegative integrable function on $\mathbb{R}$ such that $\int K(x)dx = 1$, we consider the following assumptions

$$\begin{cases} \forall x \in \mathbb{R}, K(-x) = K(x) \\ \mathrm{Supp}(K) \in [-1, 1] \\ \|K\|_\infty < \infty \end{cases} \tag{A.K}$$

We will also assume that, if $K_{h,x}^S : u \mapsto \frac{1}{h^{|S|}} \prod_{k \in S} K\left(\frac{x_k - u_k}{h}\right)$, the family of function

$$\mathcal{F}_S = \left\{K_{h,x}^S, h > 0, x \in \mathbb{R}^S\right\} \tag{4.3.2}$$

is a bounded VC class of functions. It means that it exists positive numbers A and ν such that for any probability measure P over $\mathbb{R}^S$ and any $\tau \in (0, 1)$ we have

$$\mathcal{N}\left(\mathcal{F}_S, L_2(P), \tau\right) \leq \left(\frac{A\|K\|_\infty}{\tau}\right)^\nu \tag{4.3.3}$$

where $\mathcal{N}(\mathcal{F}_S, L_2(P), \tau)$ is the τ-covering number of $\mathcal{F}_S$ for the $L_2(P)$ distance. As proved in [26] this condition is met for almost all classical kernels. An example of kernel function K satisfying all the assumptions is the Epanechnikov kernel K_{Epa}

$$K_{\text{Epa}}(x) = \frac{3}{4}(1 - x^2)\mathbb{1}_{[-1,1]}(x). \tag{4.3.4}$$

Boundary issue We must be aware of one difficulty induced by the fact that f_S is supported on $[0,1]^S$. We define the usual kernel density estimator (KDE) as follows. Let h be a positive real number. The KDE for the marginal density f_S associated with the kernel K, the bandwidth $h > 0$, and the sample W is defined as

$$\hat{f}_{h,S}^{\text{KDE}}(x) = \frac{1}{mh^{|S|}} \sum_{i=1}^{m} \prod_{k \in S} K\left(\frac{(W_i)_k - x_k}{h}\right) \tag{4.3.5}$$

We remark that even in the samples $W_1, \ldots, W_m$ belong to $[0,1]^d$, there is no reason to have $\hat{f}_{h,S}^{\text{KDE}}$ supported in $[0,1]^S$. What is more, the bias of $\hat{f}_{h,S}^{\text{KDE}}$ does not go to zero as $h \to 0$. This fact illustrates the boundary issue induced by estimating a compactly supported density.

Proposition 4.3.1

Let $f_S \in \mathcal{H}(2, L)$, the bias

$$\left\| \mathbb{E}\left[\hat{f}_{h,S}^{\text{KDE}}\right] - f_S \right\|_\infty \tag{4.3.6}$$

does not tend to 0 as $h \to 0$.

Proof.

$$\mathbb{E}\left[\hat{f}_{h,S}^{\text{KDE}}\right](0) - f_S(0) = \frac{1}{h^{|S|}} \int_{[0,1]^S} f_S(t) \prod_{k \in S} K\left(\frac{t_k}{h}\right) dt - f_S(0) \tag{4.3.7}$$

$$= \int_{[-1,1]^S} [f_S(hu) - f_S(0)] \prod_{k \in S} K(u_k) du_k \tag{4.3.8}$$

As conditions A.K are true, we have $\int xK(x) = 0$, then

$$= \int_{[-1,1]^S} \left[f_S(hu) - f_S(0) - h \sum_{k \in S} u_k \partial_k f_S(0) \right] \prod_{k \in S} K(u_k) du_k \tag{4.3.9}$$

Now as $f_S(x) = 0$ for $x \notin [0,1]^S$

$$\begin{aligned} &= \int_{[0,1]^S} \left[f_S(hu) - f_S(0) - h \sum_{k \in S} u_k \partial_k f_S(0) \right] \prod_{k \in S} K(u_k) du_k \\ &\quad - f_S(0) \int_{[-1,1]^S \setminus [0,1]^S} \prod_{k \in S} K(u_k) du_k \\ &\quad - h \sum_{k \in S} \partial_k f_S(0) \int_{[-1,1]^S \setminus [0,1]^S} u_k \prod_{k \in S} K(u_k) du_k \end{aligned} \tag{4.3.10}$$

The third term in the final sum tends to 0 with h. The same is true for the first term as

$$\begin{aligned} &\Big| \int_{[0,1]^S} \left[f_S(hu) - f_S(0) - h \sum_{k \in S} u_k \partial_k f_S(0) \right] \prod_{k \in S} K(u_k) du_k \Big| \\ &\quad \leq \int_{[0,1]^S} \left| f_S(hu) - f_S(0) - h \sum_{k \in S} u_k \partial_k f_S(0) \right| \prod_{k \in S} K(u_k) du_k && (4.3.11) \\ &\quad \leq L h^2 \sum_{k \in S} \int_{[0,1]^S} u_K^2 \prod_{k \in S} K(u_k) du_k && (4.3.12) \\ &\quad \leq L \sigma_K^2 h^2. && (4.3.13) \end{aligned}$$

Now, as $\int_{[-1,1]^S \setminus [0,1]^S} \prod_{k \in S} K(u_k) du_k = \frac{2^{|S|}-1}{2^{|S|}}$, we conclude that

$$\lim_{h \to 0} \mathbb{E}\left[\hat{f}_{h,S}^{\mathrm{KDE}} \right](0) = f_S(0) \frac{2^{|S|}-1}{2^{|S|}} \geq e^{-A|S|} \frac{2^{|S|}-1}{2^{|S|}} > 0. \tag{4.3.14}$$

□

Mirror-Image KDE To correct the boundary bias previously introduced, a solution is to add a correction to the estimator $\hat{f}_{h,S}^{\mathrm{KDE}}$ near the boundary of the

domain of definition. Let us define three mirroring operations for a number $x \in [0,1]$

$$M^{-1}(x) = -x;\ M^0(x) = x;\ M^1(x) = 2 - x. \tag{4.3.15}$$

To construct the mirror-image KDE, we start by augmenting the sample W augmented with mirror reflections of each point over all axis (see figure 4.1). Then we fit multidimensional kernels over the points of the augmented samples and restrict the domain of the obtained function to $[0,1]^S$ as illustrated in figure 4.2. Roughly speaking, it consists in flipping the part of a regular KDE constructed on the augmented dataset that falls outside $[0,1]^S$ inside it. This estimator is an extension to every dimension of the one proposed in [44]. The formal definition is

$$\hat{f}^{\mathrm{MI}}_{m,S}(x) = \mathbb{1}_{[0,1]^S}(x)\frac{1}{mh^{|S|}}\sum_{i=1}^{m}\sum_{a\in\{-1,0,1\}^S}\prod_{k\in S}K\left(\frac{M^{a_k}\left((W_i)_k\right) - x_k}{h}\right). \tag{4.3.16}$$

We remark that $\hat{f}^{\mathrm{MI}}_{m,S}$ is supported on $[0,1]^S$. The proposition 4.3.2 shows that it is a proper density, in the sense that its integral is equal to one.

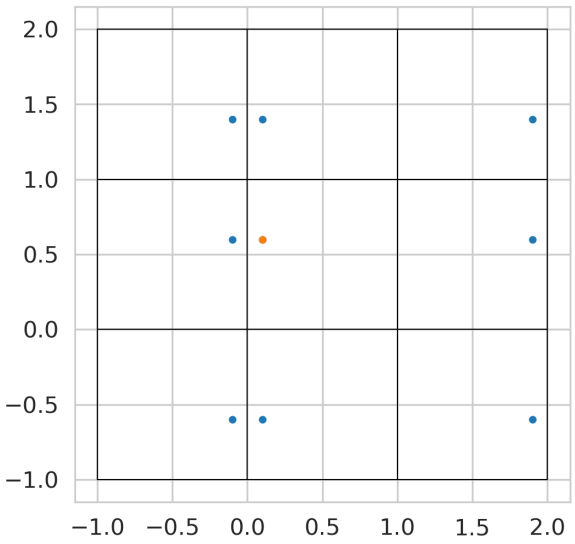
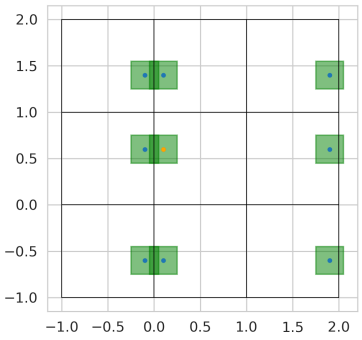
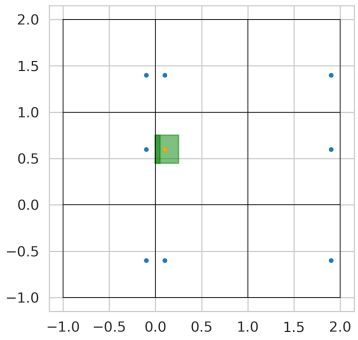


Figure 4.1: A data-point in $[0, 1]^2$ (in orange) and his 8 mirror-images (in blue)



(a) A kernel is fitted over all points of the augmented dataset



(b) We keep the restriction to the unit hypercube

Figure 4.2: Construction of the mirror-image KDE. The green areas represent the support of the kernels and the darker ones the regions where several kernels overlap (no more than two in this figure).

Proposition 4.3.2

$$\int \hat{f}^{\mathrm{MI}}_{m,S}(x)dx = 1 \tag{4.3.17}$$

Proof.

$$\int \hat{f}^{\mathrm{MI}}_{m,S}(x)dx = \int \mathbb{1}_{[0,1]^S}(x)\frac{1}{mh^{|S|}}\sum_{i=1}^{m}\sum_{a\in\{-1,0,1\}^S}\prod_{k\in S}K\left(\frac{M^{a_k}\left((W_i)_k\right)-x_k}{h}\right)dx \tag{4.3.18}$$

$$= \frac{1}{m}\sum_{i=1}^{m}\int \mathbb{1}_{[0,1]^S}(x)\sum_{a\in\{-1,0,1\}^S}\frac{1}{h^{|S|}}\prod_{k\in S}K\left(\frac{M^{a_k}\left((W_i)_k\right)-x_k}{h}\right)dx. \tag{4.3.19}$$

It is then sufficient to prove that for all $y \in [0,1]^S$

$$I_y := \int \mathbb{1}_{[0,1]^S}(x)\sum_{a\in\{-1,0,1\}^S}\frac{1}{h^{|S|}}\prod_{k\in S}K\left(\frac{M^{a_k}(y_k)-x_k}{h}\right)dx = 1. \tag{4.3.20}$$

Let us denote by $\mathcal{A}_y$ the set $\{k \in S : y_k < h \text{ or } y_k > 1-h\}$ and consider the function $K^S_{h,y} : x \mapsto \frac{1}{h^{|S|}}\prod_{k\in S}K\left(\frac{y_k-x_k}{h}\right)$. The cardinal of $\mathcal{A}_y$ corresponds to the number of direction for which the support of $K^S_{h,y}$ falls outside $[0,1]^S$. If $k \in \mathcal{A}_y$, the addition of a mirror-image of the form $x \mapsto \frac{1}{h^{|S|}}K\left(\frac{-y_k-x_k}{h}\right)$ if $y_k < h$ or $x \mapsto \frac{1}{h^{|S|}}K\left(\frac{(2-y_k)-x_k}{h}\right)$ if $y > 1-h$ will counterbalance the loss of mass induced by the restriction to $[0,1]^S$ and imply the validity of (4.3.20).

One can prove this result formally by induction. Let us consider the property P_n defined for all integer n such that $0 \le n \le |S|$ by

$$P_n : \forall y \in [0,1]^S \text{ such that } |\mathcal{A}_y| \le n, I_y = 1. \tag{4.3.21}$$

To prove P_0, let us consider that $y \in [h, 1-h]^S$, which is equivalent to assume that $|\mathcal{A}_y| = 0$. For all $k \in S$ and $x_k \in [0,1]$, we have $-y_k \le -h$ and $-x_k \le 0$. then $\frac{-y_k-x_k}{h} \le -1$. And, as the support of K is included in $[-1,1]$, we have that

$K\left(\frac{-y_k-x_k}{h}\right)=0$. Similarly, as $2-y_k \geq h+1$ and $-x_k \geq 0$, we have $\frac{(2-y_k)-x_k}{h} \geq 1$. then $K\left(\frac{(2-y_k)-x_k}{h}\right)=0$. Then

$$I_y = \int \mathbb{1}_{[0,1]^S}(x) \frac{1}{h^{|S|}} \prod_{k\in S} K\left(\frac{y_k - x_k}{h}\right) dx. \tag{4.3.22}$$

Now, as for all $k \in S$, Supp$K\left(\frac{y_k-\cdot}{h}\right) \subset [y_k - h, y_k + h] \subset [0,1]$, we have that

$$I_y = \int_{R^d} \frac{1}{h^{|S|}} \prod_{k\in S} K\left(\frac{y_k - x_k}{h}\right) dx \tag{4.3.23}$$

$$= 1. \tag{4.3.24}$$

Now, assume that P_n is true for some $n < |S|$ and let us prove that P_{n+1} is also true. Let $y \in [0,1]^S$ such that $|\mathcal{A}_y| = n+1$, and $k_0 \in S$ such that $k_0 \in \mathcal{A}_y$. We assume that $y_{k_0} < h$, the case $y_{k_0} > 1-h$ can be treated with similar arguments.

For all $x_{k_0} \in [0,1]$, as $y_{k_0} < 1-h$ we have $\frac{(2-y_{k_0})-k_{k_0}}{h} \geq 1$ and then $K\left(\frac{M^1(y_{k_0})-x_{k_0}}{h}\right)=0$. So we can decompose I_y as follows

$$I_y = \int \mathbb{1}_{[0,1]^S}(x) \sum_{a\in\{-1,0,1\}^S, a_{k_0}=-1} \frac{1}{h^{|S|}} K\left(\frac{M^{-1}(y_{k_0}) - x_{k_0}}{h}\right) \prod_{k\in S, k\neq k_0} K\left(\frac{M^{a_k}(y_k) - x_k}{h}\right) dx \tag{4.3.25}$$

$$+ \int \mathbb{1}_{[0,1]^S}(x) \sum_{a\in\{-1,0,1\}^S, a_{k_0}=0} \frac{1}{h^{|S|}} K\left(\frac{M^{0}(y_{k_0}) - x_{k_0}}{h}\right) \prod_{k\in S, k\neq k_0} K\left(\frac{M^{a_k}(y_k) - x_k}{h}\right) dx \tag{4.3.26}$$

and, using Fubini's theorem

$$I_y = \left(\int_0^1 \frac{1}{h} K\left(\frac{-y_{k_0} - x_{k_0}}{h}\right) dx_{k_0}\right) \times \lambda \tag{4.3.27}$$

$$+ \left(\int_0^1 \frac{1}{h} K\left(\frac{y_{k_0} - x_{k_0}}{h}\right) dx_{k_0}\right) \times \lambda \tag{4.3.28}$$

$$I_y = \left(\int_0^1 \frac{1}{h} K\left(\frac{-y_{k_0} - x_{k_0}}{h}\right) dx_{k_0} + \int_0^1 \frac{1}{h} K\left(\frac{y_{k_0} - x_{k_0}}{h}\right) dx_{k_0}\right) \times \lambda \tag{4.3.29}$$

where

$$\lambda = \int \mathbb{1}_{[0,1]^{S\setminus\{k_0\}}}(x) \sum_{a\in\{-1,0,1\}^{S\setminus\{k_0\}}} \frac{1}{h^{|S|-1}} \prod_{k\in S\setminus\{k_0\}} K\left(\frac{M^{a_k}(y_k)-x_k}{h}\right) dx. \quad (4.3.30)$$

Now, let us prove that $\lambda = 1$. We start by remarking that as $1/2 \in [h, 1-h]$, by the same arguments as in the proof of P_0, we have that for all $x \in [0,1]$, $K\left(\frac{M^{-1}(1/2)-x}{h}\right) = K\left(\frac{M^{1}(1/2)-x}{h}\right) = 0$. We also have $\int_0^1 \frac{1}{h} K\left(\frac{1/2-x}{h}\right) dx = 1$.

We consider $\bar{y} \in [0,1]^S$ such that $\bar{y}_k = y_k$ for all $k \neq k_0$ and $\bar{y}_{k_0} = 1/2$.

$$\lambda = \lambda \times \int_0^1 \frac{1}{h} K\left(\frac{1/2-x}{h}\right) dx \quad (4.3.31)$$

and, by Fubini's theorem

$$= \int \mathbb{1}_{[0,1]^S}(x) \sum_{a\in\{-1,0,1\}^{S\setminus k_0}} \frac{1}{h^{|S|}} K\left(\frac{1/2-x_{k_0}}{h}\right) \prod_{k\in S\setminus k_0} K\left(\frac{M^{a_k}(y_k)-x_k}{h}\right) dx \quad (4.3.32)$$

$$= \int \mathbb{1}_{[0,1]^S}(x) \sum_{a\in\{-1,0,1\}^{S}} \frac{1}{h^{|S|}} \prod_{k\in S} K\left(\frac{M^{a_k}(\bar{y}_k)-x_k}{h}\right) dx \quad (4.3.33)$$

$$= I_{\bar{y}}. \quad (4.3.34)$$

Now, as $|\mathcal{A}_{\bar{y}}| = |\mathcal{A}_y| - 1 = n$, using P_n we have $I_{\bar{y}} = 1$.

To prove that P_{n+1} is true, it remains to prove that

$$\int_0^1 \frac{1}{h} K\left(\frac{-y_{k_0}-x_{k_0}}{h}\right) dx_{k_0} + \int_0^1 \frac{1}{h} K\left(\frac{y_{k_0}-x_{k_0}}{h}\right) dx_{k_0} = 1. \quad (4.3.35)$$

In the first term of the sum, we apply the change of variable $u = -x_{k_0}$

$$\int_0^1 \frac{1}{h} K\left(\frac{-y_{k_0}-x_{k_0}}{h}\right) dx_{k_0} = -\int_0^{-1} \frac{1}{h} K\left(\frac{u-y_{k_0}}{h}\right) du \quad (4.3.36)$$

and, as K is symmetric

$$= \int_{-1}^{0} \frac{1}{h} K\left(\frac{y_{k_0} - u}{h}\right) du. \tag{4.3.37}$$

Then we have

$$\int_0^1 \frac{1}{h} K\left(\frac{-y_{k_0} - x_{k_0}}{h}\right) dx_{k_0} + \int_0^1 \frac{1}{h} K\left(\frac{y_{k_0} - x_{k_0}}{h}\right) dx_{k_0} = \int_{-1}^1 \frac{1}{h} K\left(\frac{y_{k_0} - u}{h}\right) du. \tag{4.3.38}$$

Now, as $0 \leq y_{k_0} \leq h$, we have $\text{Supp} K\left(\frac{y_{k_0} - \cdot}{h}\right) \subset [-h, 2h] \subset [-1, 1]$. Then

$$\int_{-1}^1 \frac{1}{h} K\left(\frac{y_{k_0} - u}{h}\right) du = \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{y_{k_0} - u}{h}\right) du = 1. \tag{4.3.39}$$

Then P_{n+1} is true and the proof is complete. □

Bias for mirror-image KDE Under an ad hoc condition on the partial derivatives of f_S at the boundary of $[0,1]^S$ it is possible to bound the bias for the mirror-image KDE. Our result is an extension of the lemma 3.1 in [44] to every dimension and every $\beta \in (0, 2]$ while the analysis in the original paper was restricted to bi-dimensional densities and $\beta = 2$. With our proof strategy, we find a better constant in the upper-bound for $|S| = 2$ and $\beta = 2$.

Proposition 4.3.3

Let us assume that for all sequence $(x_n)_{n\in\mathbb{N}}$ in $[0,1]^S$, if x_n converges to a boundary point of $[0,1]^S$, for all $k \in S$ we have $\lim_{n\to\infty} \partial_k f_S(x_n) = 0$. Then

$$\left\| f_S - \mathbb{E}\left[\hat{f}^{\text{MI}}_{m,S}\right] \right\|_\infty \leq C_1 h^\beta \tag{4.3.40}$$

where $C_1 = L|S|^{\beta/2} \left(2\|K\|_\infty\right)^{|S|}$ if $\beta < 2$ and $C_1 = L|S|$ if $\beta = 2$.

Proof. We define f_S^{MI} as the function defined over $[-1, 2]^S$ such that for all $x \in [0,1]^S$ and $a \in \{-1, 0, 1\}^S$

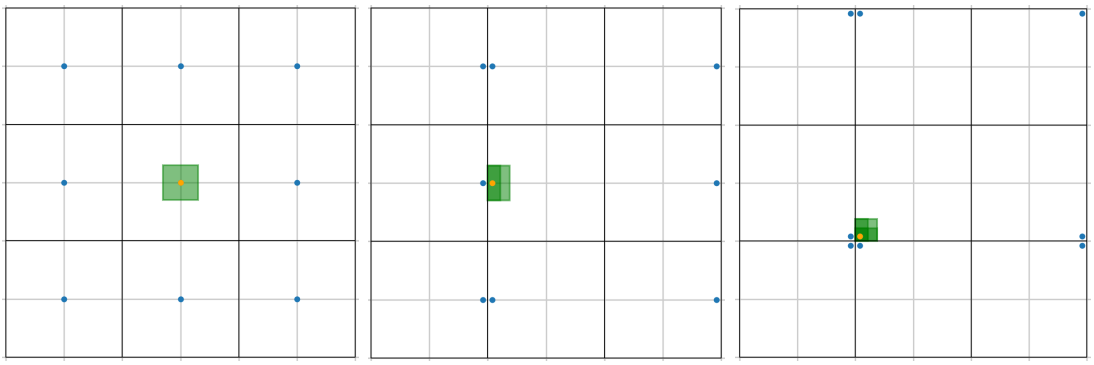
$$f_S^{\text{MI}}(M^a(x)) = f_S(X) \tag{4.3.41}$$

where $M^a(x) = (M^{a_k}(x_k))_{k\in S}$. The property that the partial derivatives of f_S vanish near the boundary of $[0,1]^S$ ensures that $\partial_k f_S^{\mathrm{MI}}$ is continuous on $(-1,2)^S$ and so $f_S^{\mathrm{MI}} \in \mathcal{H}(2,L)$.

Let $x \in [0,1]^S$, we want to bound $\left|f_S(x) - \mathbb{E}\left[\hat{f}^{\mathrm{MI}}_{m,S}(x)\right]\right|$. Assume first that $x \in [0,1/2]^S$ and denote by $\mathcal{A}$ the set $\{k \in S : x_k < h\}$. We start by considering the situation where $|\mathcal{A}| \geq 1$. For all $k \in \mathcal{A}$ and all $t \in [0,1]$, $K\left(\frac{t-(2-x_k)}{h}\right) = 0$ because the support of K is $[-1,1]$, $h \leq 1/2$ and $x_k < h$. For all $k \in S \setminus \mathcal{A}$ and all $t \in (0,1]$, $K\left(\frac{t-(2-x_k)}{h}\right) = 0$ and $K\left(\frac{t-(-x_k)}{h}\right) = 0$. Then the expected value of $\hat{f}^{\mathrm{MI}}_{m,S}$ at the point x can be written as

$$\mathbb{E}\left[\hat{f}^{\mathrm{MI}}_{m,S}\right](x) = \sum_{\mathcal{B}\subset\mathcal{A}} \frac{1}{h^{|S|}} \int_{[0,1]^S} \prod_{k\in\mathcal{B}} K\left(\frac{t_k + x_k}{h}\right) \prod_{k\in S\setminus\mathcal{B}} K\left(\frac{t_k - x_k}{h}\right) f_S(t)dt \tag{4.3.42}$$

where the sum includes a term corresponding to $\mathcal{B} = \emptyset$. For $\mathcal{B} \subset \mathcal{A}$, we denote $x_{\mathcal{B}}$ the vector such that $(x_{\mathcal{B}})_k = x_k$ if $k \notin \mathcal{B}$ and $(x_{\mathcal{B}})_k = -x_k$ if $k \in \mathcal{B}$. The indexation $\mathcal{B} \subset \mathcal{A}$ corresponds to the kernels centered in mirror-image points $(x_{\mathcal{B}})_{\mathcal{B}\subset\mathcal{A}}$ of x, for which the intersection of their support and the cube $[0,1]^S$ is nonempty. An illustration in dimension 2 is provided by figure 4.3.



(a) $h \leq x_1 \leq 1-h$, $h \leq x_2 \leq 1-h$ (b) $x_1 < h$, $h \leq x_2 \leq 1-h$ (c) $x_1 < h$, $x_2 < h$

Figure 4.3: Three examples in dimension 2 where the cardinal of $\mathcal{A}$ is zero in (a), one in (b) and two in (c). As in figure 4.2, the darker is the green area, the greater is the number of overlapping kernels. In (a), we have no overlapping, in (b) we have an overlapping of up to two kernels and in (c) an overlapping of up to four kernels.

With these notations, we can rewrite the previous formula as follows

$$\mathbb{E}\left[\hat{f}^{\mathrm{MI}}_{m,S}\right](x) = \sum_{\mathcal{B}\subset\mathcal{A}} \int_{\chi^S_{\mathcal{B}}} \prod_{k\in S} K(u_k)\, f_S(x_{\mathcal{B}} + uh) du \tag{4.3.43}$$

where $\chi^S_{\mathcal{B}} = \left\{u \in [-1,1)^S : x_{\mathcal{B}} + uh \in [0,1]\right\}$. We see that $\chi^S_{\mathcal{B}} = \prod_{k\in S}[\underline{u}_k, \bar{u}_k)$ where $\underline{u}_k = -x_k/h$ if $k \in \mathcal{B}$, -1 otherwise and $\bar{u}_k = -x_k/h$ if $k \in \mathcal{A}\setminus\mathcal{B}$, 1 otherwise. What is more, as $f^{\mathrm{MI}}_S = f_S$ on $[0,1]^S$ we have

$$\mathbb{E}\left[\hat{f}^{\mathrm{MI}}_{m,S}\right](x) = \sum_{\mathcal{B}\subset\mathcal{A}} \int_{\chi^S_{\mathcal{B}}} \prod_{k\in S} K(u_k)\, f^{\mathrm{MI}}_S(x_{\mathcal{B}} + uh) du. \tag{4.3.44}$$

Now, as $(\chi^S_{\mathcal{B}})_{\mathcal{B}\subset\mathcal{A}}$ forms a partition of $[-1,1)^S$, we have

$$f_S(x) = \sum_{\mathcal{B}\subset\mathcal{A}} \int_{\chi^S_{\mathcal{B}}} \prod_{k\in S} K(u_k)\, f_S(x) du \tag{4.3.45}$$

$$= \sum_{\mathcal{B}\subset\mathcal{A}} \int_{\chi^S_{\mathcal{B}}} \prod_{k\in S} K(u_k)\, f^{\mathrm{MI}}_S(x_{\mathcal{B}}) du \tag{4.3.46}$$

We denote by $\delta_{\mathcal{B}}(u,\beta)$ the quantity

$$\begin{cases} f^{\mathrm{MI}}_S(x_{\mathcal{B}} + uh) - f^{\mathrm{MI}}_S(x_{\mathcal{B}}) & \text{if } \beta \in (0,1] \\ f^{\mathrm{MI}}_S(x_{\mathcal{B}} + uh) - f^{\mathrm{MI}}_S(x_{\mathcal{B}}) - h\sum_{k\in S} u_k \partial_k f^{\mathrm{MI}}_S(x_{\mathcal{B}}) & \text{if } \beta \in (1,2] \end{cases} \tag{4.3.47}$$

From lemma 4.2.1 we have

$$|\delta_{\mathcal{B}}(u,\beta)| \leq L h^\beta \|u\|^\beta \tag{4.3.48}$$

And, as $\int xK(x) = 0$, we have

$$\left| f_S(x) - \mathbb{E}\left[\hat{f}^{\mathrm{MI}}_{m,S}\right](x)\right| = \left| \sum_{\mathcal{B}\subset\mathcal{A}} \int_{\chi^S_{\mathcal{B}}} \prod_{k\in S} K(u_k)\, \delta_{\mathcal{B}}(u) du \right| \tag{4.3.49}$$

$$\leq L h^\beta \int_{[-1,1]^S} \prod_{k\in S} K(u_k) \|u\|^\beta du \tag{4.3.50}$$

If $\beta = 2$

$$\int_{[-1,1]^S} \prod_{k\in S} K(u_k)\|u\|^\beta du = \int_{[-1,1]^S} \prod_{k\in S} K(u_k) \sum_{k\in S} u_k^2 du \tag{4.3.51}$$

$$= \sum_{k\in S} \int_{-1}^{1} K(u)u^2 du \tag{4.3.52}$$

$$\leq \sum_{k\in S} \int_{-1}^{1} K(u)du = |S|. \tag{4.3.53}$$

If $\beta < 2$

$$\int_{[-1,1]^S} \prod_{k\in S} K(u_k)\|u\|^\beta du \leq \|K\|_\infty^{|S|} \int_{[-1,1]^S} \|u\|^\beta du \tag{4.3.54}$$

$$\leq \|K\|_\infty^{|S|} \sqrt{|S|}^\beta \int_{[-1,1]^S} du \tag{4.3.55}$$

$$= |S|^{\beta/2} \left(2\|K\|_\infty\right)^{|S|}. \tag{4.3.56}$$

Then $\sup_{x\in[0,1/2]^S} \left| f_S(x) - \mathbb{E}\left[\hat{f}_{m,S}^{\mathrm{MI}}\right](x)\right| \leq C_1 h^\beta$. By symmetry, the same inequality is true when the sup is taken over $[0,1]^S$.

□

Then considering the mirror-image KDE leads to a correction of the boundary issue previously mentioned as the bias goes to zero when h goes to zero. What is more, the speed of convergence h^β is the usual rate for the bias of the KDE for Hölder-regular densities (see chapter 1 of [77]). The usage of mirror-image KDE only influences the constant.

Variance upper-bound

To upper-bound the variance of the mirror-image KDE, we will use corollary 15 of [35]. Our setting is not the same as in this chapter as we deal with mirror-image

KDE. Then in order to obtain the same result, we must ensure that the family of functions

$$\mathcal{F}_S^{\mathrm{MI}} = \left\{ K_{x,h}^{\mathrm{MI}} | x \in [0,1]^S, h \in (0,1/2) \right\} \tag{4.3.57}$$

where

$$K_{x,h}^{\mathrm{MI}} : u \mapsto \frac{1}{h^{|S|}} \mathbb{1}_{[0,1]^S}(u) \sum_{a \in \{-1,0,1\}^S} \prod_{k \in S} K\left(\frac{M^{a_k}(u_k) - x_k}{h} \right) \tag{4.3.58}$$

is a bounded VC class of function. We know that $\mathcal{F}_S$ is a bounded VC class of function. The results of section 2.6 of [79] indicate that a family of functions is a bounded VC class if and only if the associated collection of sub-levels is a VC class of sets. Now, we remark that the sub-levels of functions in $\mathcal{F}_S^{\mathrm{MI}}$ can be written as intersections of sub-levels of functions in $\mathcal{F}_S$ intersected with $[0,1]^S$. Then, as intersections preserve the VC class property for collection of sets (see [78]), $\mathcal{F}_S^{\mathrm{MI}}$ is a bounded VC class of functions, and the corollary 15 of [35] applies, leading to the following result.

Proposition 4.3.4

Let $h_{m,S}$ be a bandwidth in $(0,1/2)$ and $\delta_m \in (0,1)$. With probability $1-\delta_m$

$$\left\| \hat{f}_{S,h_{m,S}} - \mathbb{E}\left[\hat{f}_{m,h_{m,S}} \right] \right\|_\infty \leq C_2 \sqrt{\frac{\log\left(1/h_{m,S}\right) + \log(2/\delta_m)}{m h_{m,S}^{|S|}}}. \tag{4.3.59}$$

The constant C_2 depends on $|S|$, on $\|K\|_\infty$ and on $\|K'\|_\infty$.

Conclusion

Now, as we have for a control of the bias and the variance term for every bandwidth $h_{m,S} \in (0,1/2)$, by choosing appropriately $h_{m,S}$ it is possible to bound $\|f_S -$

$\hat{f}_{S,h_{m,S}}\|_\infty$.

Proposition 4.3.5

Choosing $h_{m,S} \asymp (1/m)^{\frac{1}{2\beta+|S|}}$, it exists a constant C_S such that with probability at least $1-\delta_m$

$$\|f_S - \hat{f}_{S,h_{m,S}}\|_\infty \leq C_S \sqrt{\log m + 2\log(2/\delta_m)} \left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+|S|}}. \tag{4.3.60}$$

4.3.2 Uniformity over Set_d^k

We have just established a control in high probability for the quantity $\|f_S - \hat{f}_{S,h_{m,S}}\|_\infty$ for a given S. Our objective is to have such a control uniformly over Set_d^k. Applying a union-bound, we obtain the following result.

Proposition 4.3.6

Let us denote $C_k = \max_{S\in\mathrm{Set}_d^k} C_S$. We have, with probability at least $1 - S_d^k\delta_m$

$$\sup_{S\in\mathrm{Set}_d^k} \left\|\hat{f}_{S,h_m^S} - f_S\right\|_\infty \leq C_k \sqrt{\log m + 2\log(2/\delta_m)} \left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+k}}. \tag{4.3.61}$$

4.4 Main theorem

Let now m_0 be the smallest integer m such that

$$C_k \sqrt{\log m + 2\log(2/\delta_m)} \left(\frac{1}{m}\right)^{\frac{2}{4+k}} \leq e^{-A|S|}\left(1 - e^{-A|S|}\right). \tag{4.4.1}$$

If $m \geq m_0$, we know that on an event $\mathcal{A}_m^k$ of probability at least $1 - S_d^k \delta_m$

$$\|f_S - \hat{f}_{S,h_m}\|_\infty \leq C_k \sqrt{\log m + 2\log(2/\delta_m)} \left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+k}}. \tag{4.4.2}$$

Then, on $\mathcal{A}_m^k$ (UC) is satisfied with $\epsilon_S = C_k \sqrt{\log m + 2\log(2/\delta_m)} \left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+k}}$ for all $S \in \mathrm{Set}_d^k$. As a consequence, using proposition 4.2.2, we have

$$\sum_{S_* \in \mathcal{P}_*} \mathrm{KL}\left(f_{S_*} \| \hat{f}_{S_*,h_m^s}\right) \leq e^{2Ak} |\mathcal{P}_*| C_k \sqrt{\log m + 2\log(2/\delta_m)} \left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+k}}. \tag{4.4.3}$$

And, on $\mathcal{A}_m^k$, for $\delta_n \in (0, 1/S_d^k)$ with probability $1 - S_d^k \delta_n$

$$(P - P_n)(\log \hat{f}_{\tilde{\mathcal{P}}} - \log \hat{f}_{\hat{\mathcal{P}}}) \leq 4\sqrt{2} d \frac{Ak}{\sqrt{n}} \sqrt{\log(2/\delta_n)}. \tag{4.4.4}$$

Now, with the choices $\delta_m = 2/(S_d^k m)$ and $\delta_n = 2/(S_d^k n)$ we obtain the following result

Theorem 4.4.1

If for all $S \in \mathrm{Set}_d^k$, $f_S \in \mathcal{H}(\beta, L)$, satisfies BC, and for all sequence $(x_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} x_n = x^*$ where x^* belongs to the boundary of $[0,1]^S$ and for all $k \in S$ $\lim_{n\to\infty} \partial_k f_S(x_n) = 0$. With the choice $\hat{f}_S = \hat{f}^{\mathrm{MI}}_{h_{m,S},S}$ where $h_{m,S} \approx \left(\frac{1}{m}\right)^{\frac{1}{2+|S|}}$, we have with probability at least $(1 - 2/m)(1 - 2/n)$

$$\begin{aligned} \mathrm{KL}\left(f \| \hat{f}_{\hat{\mathcal{P}}}\right) - \mathrm{KL}\left(f \| f_{\mathcal{P}_*}\right) \leq & e^{2Ak}\sqrt{2}|\mathcal{P}_*| C_k \sqrt{\log m + \log\left(S_d^k\right)} \left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+k}} \\ & + 4\sqrt{3} d \sqrt{\log n + \log\left(S_d^k\right)} \frac{Ak}{\sqrt{n}} \end{aligned} \tag{4.4.5}$$

Ignoring logarithmic factors, the rate of convergence of the approximation term is $\left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+k}}$. The dependence of this quantity in k illustrates that ISDE tackles the curse of dimensionality for the density estimation problem under KL loss in the same spirit that [65] showed that an adapted estimator does for the squared L_2 loss.

Ignoring logarithmic factors again, the rate of convergence of the selection term is $\frac{1}{\sqrt{n}}$. This is a classical rate of convergence for hold-out procedures with bounded loss (see corollary 8.8 in [50]). We can ask whether it could be possible to obtain a faster rate of convergence for this term. Indeed, theorem 8.9 of [50] indicates that we can obtain a rate of convergence of $\frac{1}{n}$ by introducing the Kullback-Jensen loss $KJ(f, \hat{f}_{\hat{\mathcal{P}}}) := \mathrm{KL}\left(f \| \frac{f+\hat{f}_{\hat{\mathcal{P}}}}{2}\right)$. This loss is equivalent to KL under UC. However, the drawback of this approach is that KJ does not behave as KL with a product of densities, and the combinatorial price would be the cardinal of Part_d^k.

The term $\log(S_d^k)$ in the upper-bound illustrates the combinatorial complexity reduction operated by ISDE. The presence of the log of the number of hypotheses is classical for hold-out procedures with bounded loss (see again corollary 8.8 in [50]). In our context, we have reduced the combinatorial complexity from the number of partitions P_d^k to the number of subsets S_d^k.

4.5 Conclusion

In this chapter, we have studied the convergence properties of ISDE. In particular, we have shown that under suitable assumptions on the true density and for the mirror-image KDE as marginal density estimator, we can provide an upper-bound valid with high-probability of the quantity

$$\mathrm{KL}\left(f \| \hat{f}_{\hat{\mathcal{P}}}\right) - \mathrm{KL}\left(f \| f_{\mathcal{P}_*}\right). \tag{4.5.1}$$

This bound highlights how ISDE tackles the curse of dimensionality and reduces the combinatorial complexity of the density estimation problem under IS compared to a brute-force approach. These results offer a theoretical validation of the empirical observations presented in [62].

To complete the study, it let to study how the bias term $\mathrm{KL}\left(f \| f_{\mathcal{P}_*}\right)$ behaves. It is hard to give a precise statement on this quantity in a general setting. One simple situation is when $f \in \mathcal{D}_k^d$. In this case, $\mathrm{KL}\left(f \| f_{\mathcal{P}_*}\right) = 0$. The bias can also be explicitly evaluated in some simple frameworks, such as multivariate Gaussian variables; detailed computations are presented in appendix B. Unfortunately, the

Gaussian framework does not fill in the scope of the present chapter as it involves variables that does not respect the condition (BC).

Chapter 5

Application of ISDE to cytometry data

Context Modern cytometry experiments tend to use more and more markers. For statisticians, it means that we need to deal with a bigger value of d and one can expect that this trend will continue to be true in the next years or decades. To anticipate these changes, it is important to design techniques that are still relevant in high dimension. UMAP is an appealing solution, and it has proved to be efficient for cytometry data representation ([5], [4]) and identification of rare cell populations ([86]). However, UMAP acts as a black-box method. This is why we have decided to follow a different path and incorporate a dimensionality reduction scheme in the density estimation step via ISDE in a density-based clustering algorithm.

Our contribution After introducing and studying computational aspects of ISDE in chapter 3 and proving an oracle inequality for it in chapter 4, we demonstrate the interest of using ISDE on cytometry data. We focus on the datasets presented in the benchmark of [85]. We show that in terms of log-likelihood on validation data, ISDE outperforms other density estimation techniques for mass cytometry data. We also study the space of partitions over which ISDE optimizes and show that the topology induced by the edit distance between partitions correlates well with the decrease in the log-likelihood on validation data when going far from the partition outputted by ISDE.

Organization of the chapter In section 5.1 we compare ISDE and other density estimation methods (KDE, FDE and Gaussian Mixture) in terms of log-likelihood on validation data. In section 5.2 we study the repartition of the log-likelihoods over the space of partitions for these experiments with the topology induced by the edit distance. Then in section 5.3 we show that combining ISDE with CyTOMATo leads to an improvement of the F1 score obtained on the Weber and Robinson data in comparison to what have been obtain with CyTOMATo in chapter 2.

5.1 Quantitative evaluation

ISDE has been designed to maximize log-likelihood through the combination of a multivariate density estimator (in our experiments, we use a Gaussian KDE with cross-validated bandwidth) and the learning of a partition of the variables. As in section 3.3 we can qualitatively compare the performance of ISDE against other density estimation methods.

Benchmarked Algorithms As in section 3.3, we compare FDE, CVKDE, and ISDE (the value of k depends on the dimension, we selected $k = 3$ for Levine32, Samusik01 and SamusikAll and $k = 5$ for Levine13 to keep computations time reasonable).

We have also added a parametric approach to the benchmark: a Gaussian Mixture (GM) model with a selection of the number of components. This model is particularly adapted to cytometry as we naturally expect in this context that the data forms clusters representing cell populations ([66], [24]).

Let n_C be a positive integer corresponding to the number of components in the mixture. Let $p = (p_1, \ldots, p_{n_C})$ be a collection of nonnegative real number such that $\sum_{i=1}^{n_C} p_i = 1$, $\mu = (\mu_1, \ldots, \mu_{n_C})$ a collection of vectors in $\mathbb{R}^d$ and $\Sigma = (\Sigma_1, \ldots, \Sigma_{n_C})$ a collection of $d \times d$ definite positive matrices. The density $f_{(n_C,p,\mu,\Sigma)}$ of the Gaussian mixture model associated with the parameters (n_C, p, μ, Σ) is

$$f_{(n_C,p,\mu,\Sigma)} = \sum_{i=1}^{n_C} p_i f_{\mu_i,\Sigma_i} \tag{5.1.1}$$

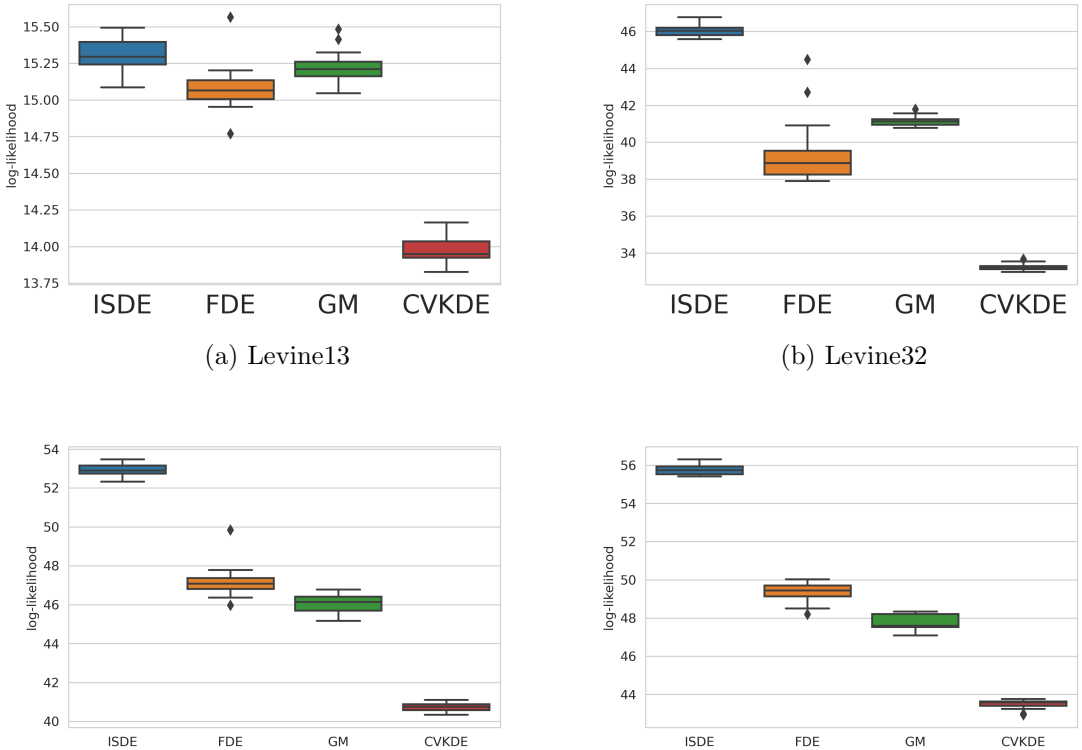
where f_{μ_i,Σ_i} is the density of the multivariate Gaussian random variable with mean μ_i and covariance matrix Σ_i.

Given n_C and a dataset, it is possible to compute estimators $(\hat{p}, \hat{\mu}, \hat{\Sigma})$ with the EM algorithm [21] to maximize the log-likelihood. As we do not know the optimal number of components in advance, a strategy is to fit a Gaussian mixture model for different n_C (from 1 to 30 in our experiments) and select the number of components in the mixture with a cross-validation scheme. We rely on the implementation of these methods provided by scikit-learn [10] with no restriction on the shape of the covariance matrices.

Though GM is principally used for clustering purposes, it can also be interpreted as a parametric density estimator intended to maximize the log-likelihood. It is then relevant to compare it with the other introduced methods.

Experimental Setup From each dataset we have extracted a train sample with $N = 5000$ events, this train sample is exclusively used to compute estimators $\hat{f}_{\text{CVKDE}}$, $\hat{f}_{\text{FDE}}$, $\hat{f}_{\text{ISDE}}$ and $\hat{f}_{\text{GM}}$. For ISDE we fixed $m = 3000$ and $n = 2000$. Then to compare between these density estimators, we sampled 20 datasets with 2000 events from the data that were not used to compute estimators.

Results Box plots indicating the log-likelihood of these estimators for validation samples can be visualized in figure 5.1.



(a) Levine13

(b) Levine32

(c) Samusik01

(d) SamusikAll

Figure 5.1: log-likelihood on test data for different density estimation methods

We remark that using ISDE leads to better empirical log-likelihood on validation data. CVKDE in the ambient dimension is always the worst estimator. GM is slightly better than FDE for Levine13 and Levine32 datasets, and the opposite is true for Samusik01 and SamusikAll. The relative gap between performances of FDE/GM and ISDE seems to be higher in high dimensions (recall that $d = 13$ for Levine13, $d = 32$ for Levine32 and $d = 39$ for Samusik01 and SamusikAll). We conclude that IS with a limited size of blocks seems to be a relevant model for these datasets, as ISDE could outperform other model-based approaches in terms of log-likelihood.

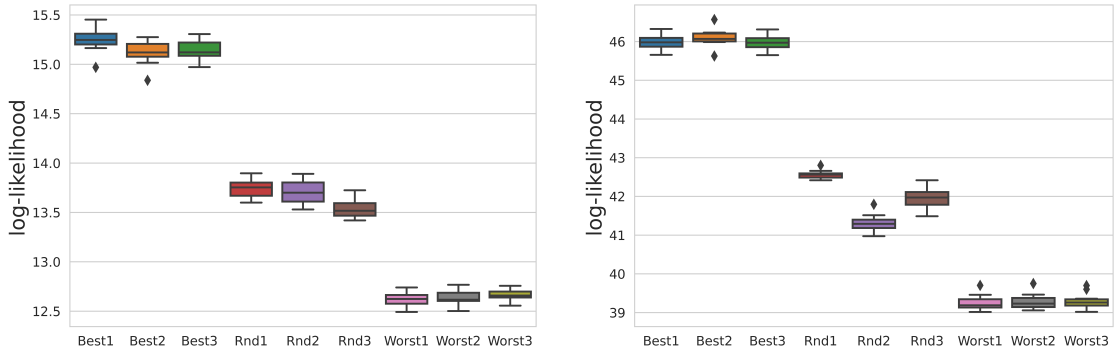
Testing ISDE against other density estimation methods is a way to evaluate how this model can explain the data well. However, we must be careful in our conclusion. These results do not indicate that the data follow an IS, but rather that IS offers a good approximation of the data distribution comparatively to other density estimation approaches.

5.2 Qualitative interpretation

We believe that the added value of our method is that ISs are easy to understand and usable as a tool to interpret data. After validating the pertinence of ISDE in comparison with other methods through quantitative analysis, we now provide some insight into the capacity of ISDE to deliver meaningful qualitative information.

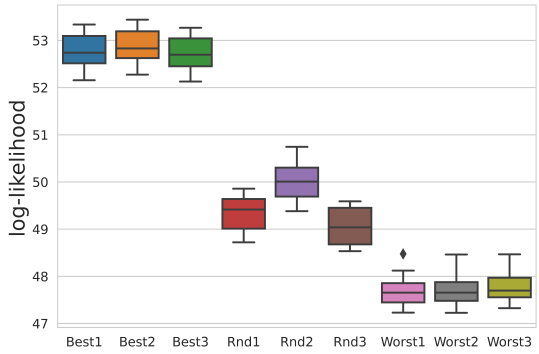
Non-triviality of Outputted Partition The first question to ask is if the gain in terms of empirical log-likelihood is due to the specific outputted partition $\hat{\mathcal{P}}$ or if any other estimator $\hat{f}_{\mathcal{P}}$ based on a partition of features $\mathcal{P} \in \mathrm{Part}_d^k$ could achieve the same performance. To answer this question, we have computed empirical log-likelihood on 10 validation sets of size $2,000$ for the three best partitions outputted by ISDE, the three worst ones regarding the optimization task, and three random partitions in Part_d^k for all of the four datasets. To compute not the optimal but the second one, the third one, and so on, it suffices to add constraints on the partition selection problem that artificially exclude some partitions from the optimization. To compute the worst partitions, switching the optimization from maximization to minimization suffices. Random partitions are computed by

generating a random permutation σ of $\{1, \ldots, d\}$ and then gather consecutive features in $\{\sigma(1), \ldots, \sigma(d)\}$ in groups with sizes drawn uniformly between 1 and k. The results are presented in figure 5.2.

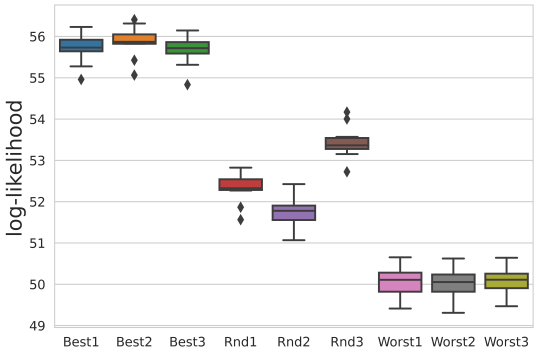


(a) Levine13

(b) Levine32



(c) Samusik01



(d) SamusikAll

Figure 5.2: log-likelihood on test data for 3 bests, 3 randoms and 3 worsts partitions

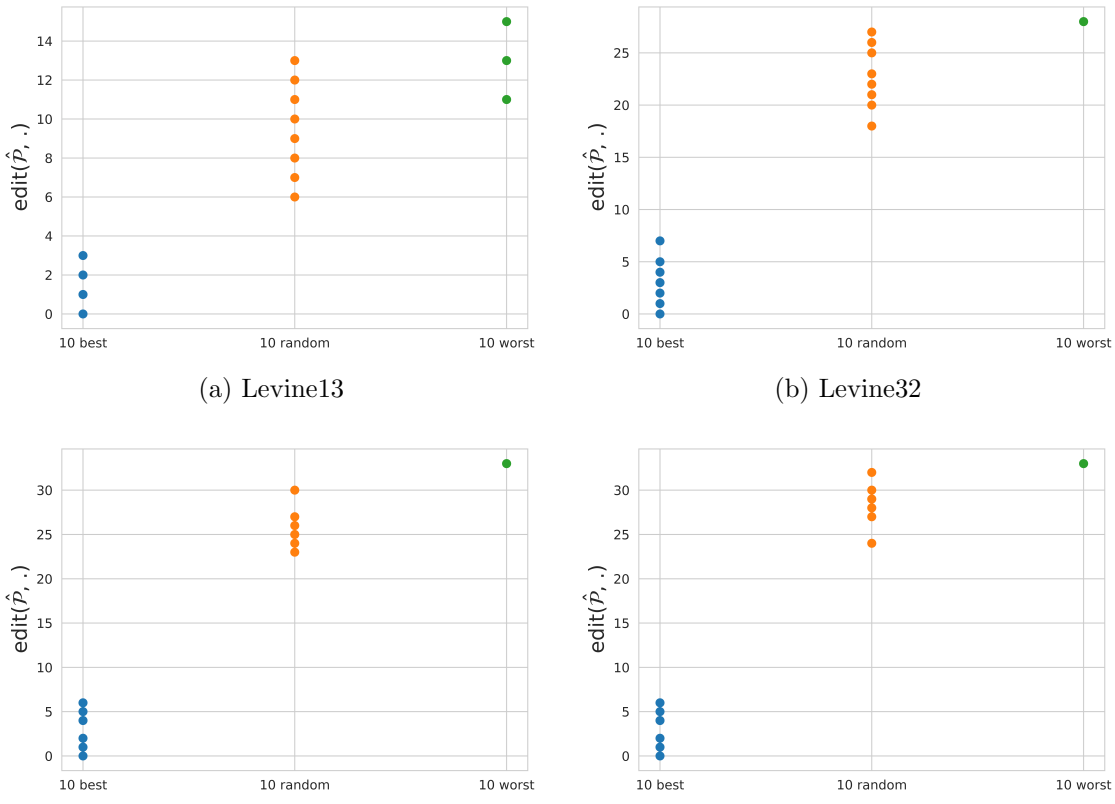
These experiments indicate that ISDE outputs specific partitions that lead to better estimators in terms of log-likelihood on empirical data than random partitions. In that sense, the information provided by ISDE on these datasets is not trivial. It also seems that not only the optimal one $\hat{\mathcal{P}}$, but a collection of partitions lead to the best scores.

With that in mind, it could be interesting to determine if the collection of partitions leading to optimal results are close in some sense. To this end, it is

necessary to introduce a notion of distance between partitions.

Edit Distance Given two partition $\mathcal{P}$ and $\mathcal{P}'$ in Part_d^k it is possible to define a distance between $\mathcal{P}$ and $\mathcal{P}'$ called edit distance ([9]) and denoted by $\text{edit}(\mathcal{P}, \mathcal{P}')$. This distance corresponds to the minimal number of operations required to go from $\mathcal{P}$ to $\mathcal{P}'$ where an operation can split a block into two or merge two blocks. The edit distance defines a distance on Part_d^k in the mathematical sense as it is nonnegative, symmetric, equal to zero only if we compute the distance from one partition to itself, and it satisfies the triangular inequality.

Correlation between Edit Distance and Density Estimation We will now see how the edit distance from $\hat{\mathcal{P}}$ to $\mathcal{P}$ correlates with the empirical log-likelihood on validation data for $\hat{f}_{\mathcal{P}}$. Firstly, we can visualize the edit distance from $\hat{\mathcal{P}}$ to the 10 best partitions (excluding $\hat{\mathcal{P}}$) in the sense of the problem of partition selection, 10 random partitions, and the 10 worst partitions. See figure 5.3.



(a) Levine13

(b) Levine32

(c) Samusik01

(d) SamusikAll

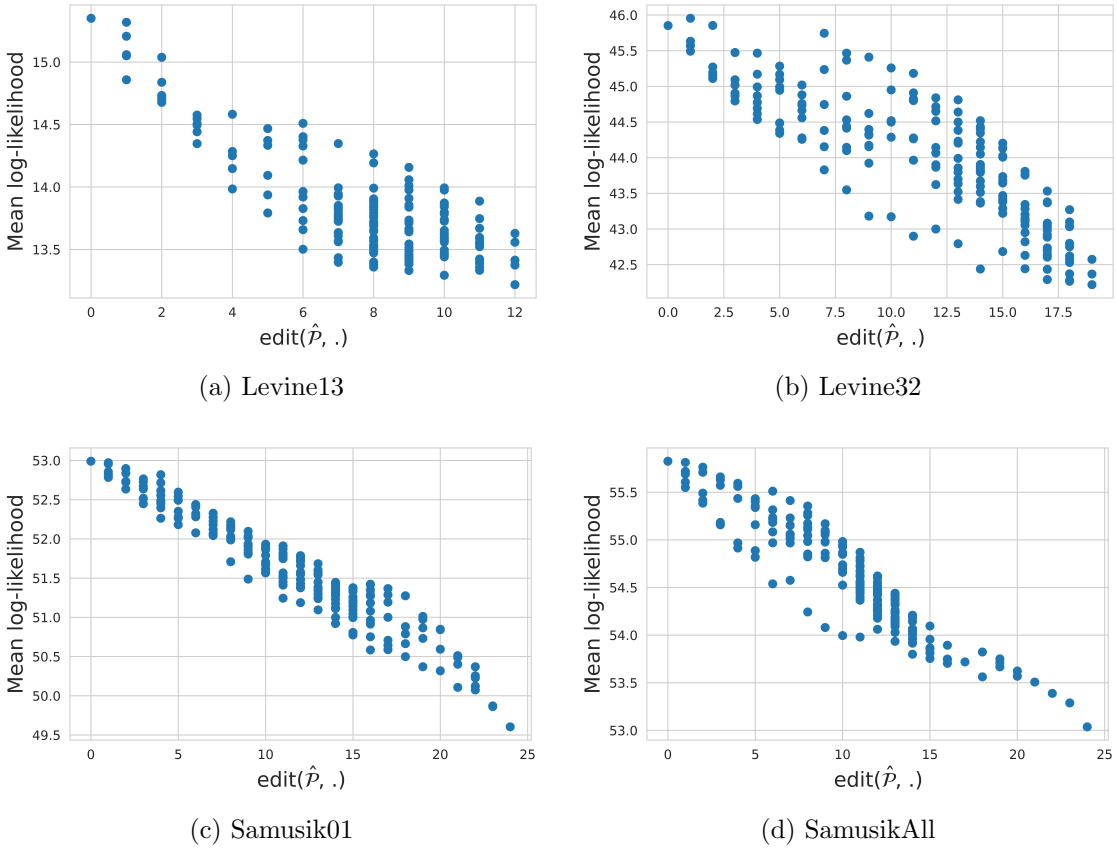
Figure 5.3: log-likelihood on test data for different density estimation methods

These observations seem to correlate well with what we have observed previously in terms of log-likelihood.

Secondly, we can explore the space Part_d^k by defining a random walk considering the topology induced by the edit distance. We define a random walk $(\mathcal{P}_0, \mathcal{P}_1, \dots)$ as follows: at each step we go from $\mathcal{P}_i$ to $\mathcal{P}_{i+1}$ with $\text{edit}(\mathcal{P}_i, \mathcal{P}_{i+1}) = 1$. To do so, it suffices to randomly choose an operation (edit or merge) and apply it to randomly selected block(s) of $\mathcal{P}_i$ while controlling that we stay in Part_d^k.

To observe a possible correlation between $\text{edit}(\hat{\mathcal{P}}, .)$ and log-likelihood on validation data, we have implemented the following protocol: do 5 walks of length 40 with $\hat{\mathcal{P}}$ as starting point and store all visited partitions, then for the 200 selected partitions, compute empirical log-likelihood on ten resampling of validation data

and store the mean value. Then we plot these scores against edit($\hat{\mathcal{P}}$, .).



(a) Levine13

(b) Levine32

(c) Samusik01

(d) SamusikAll

Figure 5.4: log-likelihood on test data with respect to edit distances

For all datasets, we observe a clear negative correlation between edit($\hat{\mathcal{P}}$, .) and empirical log-likelihood on validation data. These observations indicate that the topology induced by the distance edit on Part_d^k is meaningful in the sense that the farther a partition $\mathcal{P}$ is from $\hat{\mathcal{P}}$ for the edit distance, the worse the estimator $\hat{f}_{\mathcal{P}}$ is.

Exhaustive Analysis For the dataset Levine13, as the cardinal of Part_{13}^5 is $25,719,630$, it is possible to store the entire family of empirical log-likelihood computed thanks to the data $Z_1, \ldots, Z_n$ on ISDE: $(\ell_n(\mathcal{P}))_{\mathcal{P} \in \text{Part}_{13}^5}$. Such an exhaustive analysis is impossible for Levine32, Samusik01 and SamusikAll as the number of

partitions in Part_{32}^{3} exceed 10^{19}. For Levine13, we represented the distribution of $(\ell_n(\mathcal{P}))_{\mathcal{P}\in\mathrm{Part}_{13}^{5}}$ thanks to an histogram in figure 5.5.

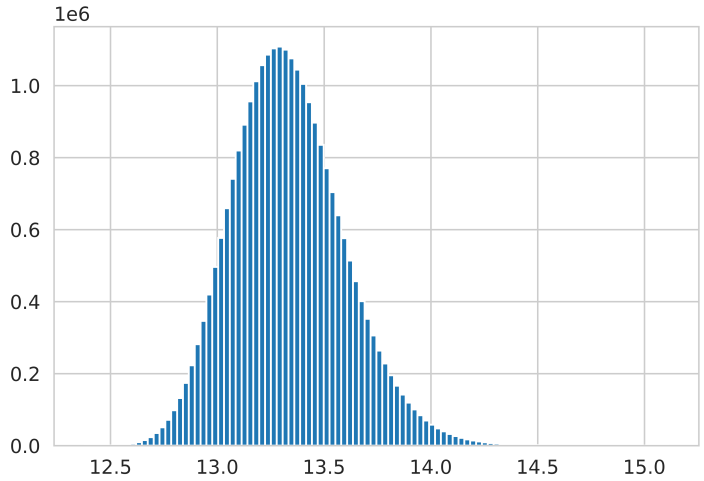


Figure 5.5: Distribution of $(\ell_n(\mathcal{P}))_{\mathcal{P}\in\mathrm{Part}_{13}^{5}}$

If we select the partitions with a score higher than 14.6, there remain $1,941$ elements. For these partitions, we compute empirical log-likelihood again on validation data and represent it against $\mathrm{edit}(\hat{\mathcal{P}},.)$. This is a way to ask about the uniqueness of the optimal partition $\hat{\mathcal{P}}$. If another partition $\mathcal{P}$ a significantly positive value of $\mathrm{edit}(\hat{\mathcal{P}},\mathcal{P})$ gives as good results as $\hat{\mathcal{P}}$, it will indicate that there are other local maximums than $\hat{\mathcal{P}}$. The outputs are represented in figure 5.6.

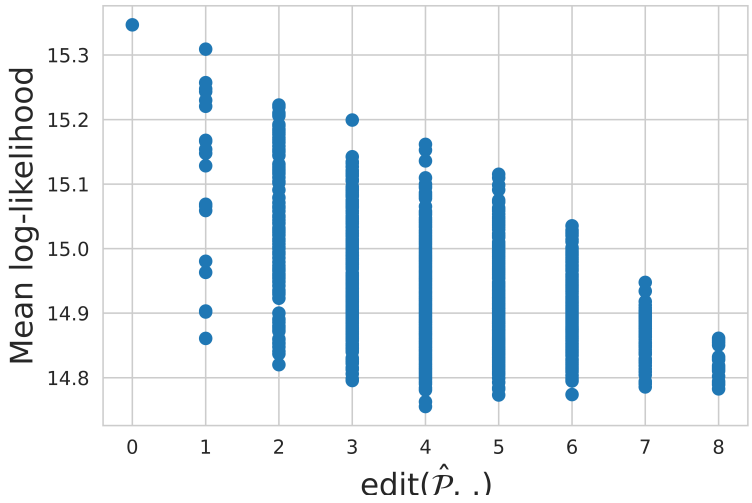


Figure 5.6: Mean log-likelihood on validation data with respect to edit distance from $\hat{\mathcal{P}}$ for $1,941$ best partitions

Conclusion This analysis of the space Part_d^k equipped with edit distance in terms of empirical log-likelihood for $\hat{f}_{\mathcal{P}}$ has led us to the conclusion that the qualitative information provided by ISDE through $\hat{\mathcal{P}}$ is nontrivial for these datasets as random partitions in Part_d^k does not lead to optimal scores. We also show that the density estimation score deteriorates as the edit distance from $\hat{\mathcal{P}}$ increases, indicating that edit distance is a relevant metric to explore Part_d^k in density estimation under IS. Then an exhaustive analysis of the space of partitions for Levine13 indicates that we can consider the optimal partition as unique for this experiment.

These conclusions depend on the specific datasets presented here and could become invalid for other ones. We provide the code to reproduce our experiments. Our aim is that anyone interested in the method can replicate these analyses for other data.

5.3 Combining ISDE density estimation with CyToMATo

One of our initial motivations for introducing ISDE was the potential improvement of the clustering obtained with ToMATo through a dimensionality reduction step incorporated in the density estimation phase. We have just proven that ISDE offers qualitative information that can be useful for the cytometrist, but can this information help us improve the quality of clustering delivered by CyToMATo ?

To answer this question, the idea is to slightly modifying CyToMATo as presented in chapter 2 by switching the density estimator from the DTM-based density estimator to an estimator taking the independence structure learned by ISDE into account. More precisely, if $\hat{\mathcal{P}}$ is the partition of variables outputted by ISDE, we define the density estimator

$$g_{\mathcal{P},20}^{\text{logDTM}} = \sum_{S \in P} g_{S,20}^{\text{logDTM}} \tag{5.3.1}$$

where $g_{S,20}^{\text{logDTM}}$ denotes the log-DTM density estimator with parameter $k_d = 20$ on $(X_1)_S, \ldots, (X_N)_S$. We use this estimator as an alternative of in the definition of CyToMATo, this alternative is denoted as CyToMATo_ISDE. One can then compare the performance of this new alternative on the data of the Weber and Robinson benchmark, as in sections 2.3 and 2.4. Note that we did not recompute ISDE and that we used the partitions obtained in section 5.1.

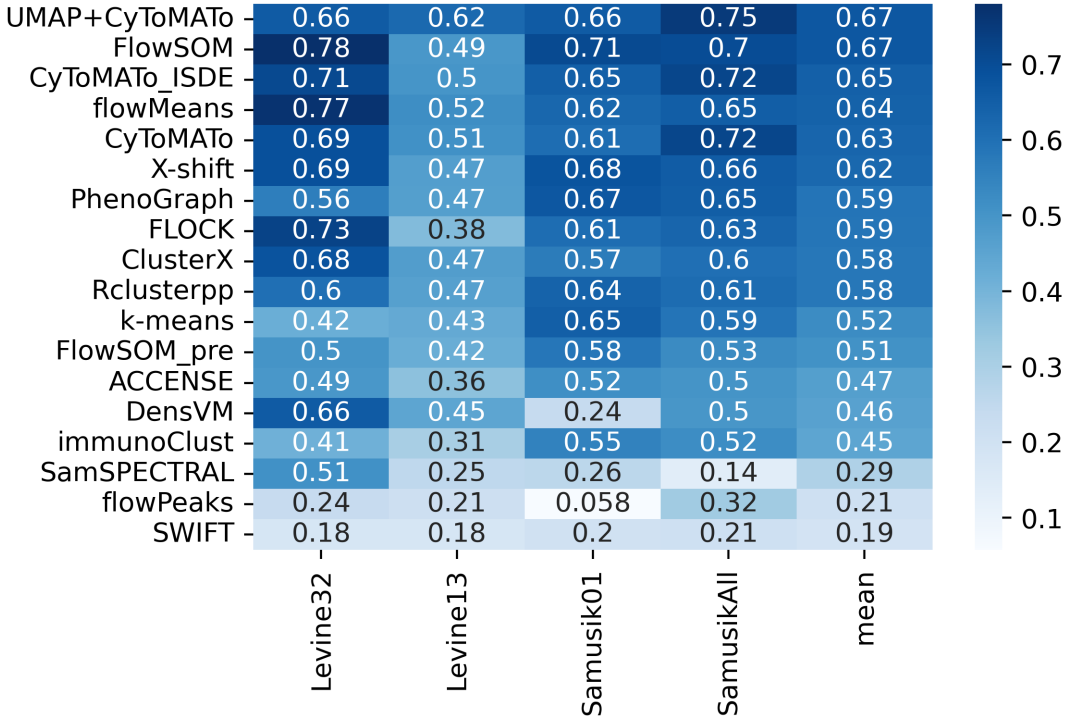


Figure 5.7: Comparison of clustering scores of CyTOMATo, and UMAP+CyToMATo, CyToMATo_ISDE and other algorithms tested in [85] for Levine13, Levine32, Samusik01 and SamusikAll and the mean of the four scores

We observe a slight improvement in terms of F1 scores comparatively to CyToMATo. However, scores remain lower than what was obtained with FlowSOM or UMAP+CyToMATo. Note that we have not considered directly the KDE estimator outputted by ISDE as input for ToMATo. As explained in section section 2.2, we observed poor clustering results using KDE estimators.

5.4 Conclusion

In this chapter, we have highlighted the pertinence of ISDE in the context of cytometry data analysis from three point of view. Firstly, we have shown that ISDE

have a good performance comparatively to other density estimation approaches, evaluated through log-likelihood on test data. Secondly, we have presented evidence that the edit distance to the partition outputted by ISDE is well correlated with the empirical log-likelihood. Thirdly, we have proposed a modification of CyTOMATo and shown that it has led to a slight improvement in terms of F1 score compared to CyToMATo.

The take-home message of this chapter is that the IS hypothesis is relevant for cytometry data and ISDE deliver an information that can be useful to improve clustering or to extract qualitative information about data through the partition of the variables.

Chapter 6

Conclusion et perspectives

6.1 Conclusion

Dans cette thèse, nous avons développé une approche originale pour le partitionnement des données de cytométrie. Le parti pris de notre démarche a été de ne pas nous limiter à chercher à améliorer un score quantitatif de comparaison à un partitionnement expert. En échangeant avec des cytométristes, dans le cadre de notre collaboration avec Metafora, nous avons acquis la conviction que l'adoption de méthodes de partitionnement automatique étaient freinée par deux aspects. D'une part, la difficulté à utiliser les approches existantes par manque d'interfaces intuitives et d'algorithmes simples à calibrer dissuade les cytométristes d'utiliser ces approches. D'autre part, l'objectif à long terme d'intégrer ces méthodes dans des contextes cliniques nécessite d'éviter le recours à des approches de type "boîte noire", cet aspect est souvent négligé, notamment lorsque des approches de réduction par des algorithmes comme UMAP sont intégrées. Nous avons donc contribué à proposer des pistes pour résoudre ces problèmes. L'algorithme CytOMATo ne nécessite pas de calibrer des hyper-paramètres et permet d'obtenir un partitionnement hiérarchique des données. Nous pensons que c'est un bon candidat pour une intégration dans une interface à destination des experts du domaine. De plus, la démarche d'estimation de densité définie par ISDE permet de tirer une information qualitative sur les données via la structure d'indépendance.

Nous avons aussi tenu à présenter l'algorithme ISDE indépendamment du cadre de la cytométrie. Nous pensons en effet que l'exploitation de l'hypothèse de structure d'indépendance sur des données de dimension modérée pourrait être pertinente dans d'autres cadres applicatifs. Il nous semble pertinent de comparer ISDE avec l'algorithme FDE précédemment développé. Ces approches présentes des similarités. Elles proposent toutes deux d'estimer des estimateurs marginaux à l'aide de noyaux (bien qu'elles puissent s'adapter avec d'autres estimateurs) puis d'apprendre un modèle graphique d'une forme particulière à partir de ces estimations. Les modèles que peuvent apprendre ces deux algorithmes sont disjoints, nous pensons donc qu'une étude comparative des deux approches est pertinente sur de nouvelles données. De plus, les estimateurs marginaux obtenus avec ISDE peuvent être réutilisés pour exécuter FDE en temps linéaire en la dimension. Pour l'implémentation d'ISDE avec des estimateurs à noyaux, nous avons bénéficié des avancées récentes dans le domaine du calcul sur carte graphique, avec en particulier l'utilisation de la librairie KeOps [12]. Les temps de calculs nous permettent de calibrer les largeurs de bandes à l'aide de validations croisées.

Enfin, nous avons proposé une majoration en grande probabilité pour la perte KL de l'estimateur en sortie d'ISDE. La borne traduit à la fois un gain par rapport au fléau de la dimension et un gain combinatoire lié à la structure de l'algorithme ISDE. Pour ce faire, nous avons eu à proposer des estimateurs originaux (en particulier une extension au cadre multidimensionnel de l'estimateur à noyaux miroir proposé dans [44] dans le cadre de la dimension 2) et à mobiliser des résultats récents d'estimation de densité en perte uniforme (notamment ceux de [35]). L'hypothèse clé est le fait que la vraie densité soit bornée inférieurement.

6.2 Perspectives

Du point de vue de l'application à la cytométrie, nous espérons qu'à l'avenir, ce travail donnera lieu à une intégration logicielle utile aux cytométristes et sera une source d'inspiration pour le développement de méthodes automatiques qui pourront à terme être utilisées dans le cadre d'applications cliniques. Nous espérons aussi que notre travail concernant la structure d'indépendance trouvera son utilité pour l'analyse descriptive des données de cytométrie.

Concernant ISDE, certaines limitations demeurent. En terme de temps de calcul, nous devons toujours limiter le paramètre k quand la dimension augmente. Nous avons fait le choix de l'exhaustivité dans l'exploration des structures et dans la sélection des largeurs de bande par validation croisée. Peut-être sera-t-il intéressant à l'avenir de réfléchir à des versions moins exhaustives mais plus rapides d'ISDE pour traiter des problème de plus grande dimension et gagner en vitesse d'exécution. Le recours à l'utilisation de calcul sur carte graphique nous a permis une accélération significative des temps de calcul des noyaux gaussiens. Ce type de calcul est à la base de nombreuses méthodes, et il est intéressant aujourd'hui de proposer de nouvelles implémentations d'algorithmes déjà existants mais qui ont été écrits alors que ces techniques n'étaient pas disponibles. En particulier, alors que l'implémentation proposée pour FDE dans [45] proposait de fixer la largeur de bande des noyaux gaussien à partir de la taille de l'échantillon, il devient aujourd'hui possible de remplacer cette étape par une validation croisée dès lors que N reste de l'ordre de quelques milliers.

Du point de vue théorique, le résultat du chapitre 4 laisse des questions ouvertes. On peut par exemple se demander si l'obtention d'une borne en probabilité et non en espérance est liée à une limitation intrinsèque au modèle ou si une autre technique de preuve aurait mené à un résultat en espérance. On peut aussi s'interroger sur l'optimalité de la borne en $N^{-\frac{\beta}{2\beta+k}}$. Sous les hypothèses de bornitudes que nous considérons sur la densité et sur les estimateurs, on a équivalence entre la perte de Kullback-Leibler et la perte quadratique au carré. on s'attendrait alors à obtenir une majoration du risque en $N^{-\frac{2\beta}{2\beta+k}}$ en accord avec les résultats de Rebelles [65]. Est-ce ici une faiblesse de notre approche ou sommes nous intrinsèquement limités par le fait que nous n'utilisons pas la connaissance des bornes de la densité f dans la construction des estimateurs ? Enfin, on pourrait chercher à améliorer notre résultat en le rendant adaptatif à la régularité β et au paramètre k, et en obtenant une vitesse rapide pour le terme de sélection. Ces questions demeurent ouvertes au moment de la rédaction de ce manuscrit, nous espérons y apporter des réponses prochainement.

Bibliography

[1] Nima Aghaeepour et al. "Rapid cell population identification in flow cytometry data". In: *Cytometry Part A* 79.1 (2011), pp. 6–13.

[2] Sylvain Arlot and Alain Celisse. "A survey of cross-validation procedures for model selection". In: *Statistics surveys* 4 (2010), pp. 40–79.

[3] Dmitry R Bandura et al. "Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry". In: *Analytical chemistry* 81.16 (2009), pp. 6813–6822.

[4] Etienne Becht et al. "Dimensionality reduction for visualizing single-cell data using UMAP". In: *Nature biotechnology* 37.1 (2019), pp. 38–44.

[5] Etienne Becht et al. "Evaluation of UMAP as an alternative to t-SNE for single-cell data". In: *BioRxiv* (2018), p. 298430.

[6] Gérard Biau et al. "A weighted k-nearest neighbor density estimate for geometric inference". In: *Electronic Journal of Statistics* 5 (2011), pp. 204–237.

[7] C Blanc et al. "Intérêt de la numération absolue par cytométrie en flux et du quadruple marquage des sous-populations lymphocytaires lors de l'infection par le VIH". In: *Revue française des laboratoires* 1996.287 (1996), pp. 59–64.

[8] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[9] Duncan P Brown, Nandini Krishnamurthy, and Kimmen Sjölander. "Automated protein subfamily identification and classification". In: *PLoS computational biology* 3.8 (2007), e160.

[10] Lars Buitinck et al. "API design for machine learning software: experiences from the scikit-learn project". In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 2013, pp. 108–122.

[11] Danielle L Cantrell et al. "The use of kernel density estimation with a biophysical model provides a method to quantify connectivity among salmon farms: spatial planning and management with epidemiological relevance". In: *Frontiers in Veterinary Science* (2018), p. 269.

[12] Benjamin Charlier et al. "Kernel Operations on the GPU, with Autodiff, without Memory Overflows". In: *Journal of Machine Learning Research* 22.74 (2021), pp. 1–6. URL: `http://jmlr.org/papers/v22/20-275.html`.

[13] Frédéric Chazal and Bertrand Michel. "An introduction to topological data analysis: fundamental and practical aspects for data scientists". In: *arXiv preprint arXiv:1710.04019* (2017).

[14] Frédéric Chazal et al. "Persistence-based clustering in Riemannian manifolds". In: *Journal of the ACM (JACM)* 60.6 (2013), pp. 1–38.

[15] Frédéric Chazal et al. "Proximity of persistence modules and their diagrams". In: *Proceedings of the twenty-fifth annual symposium on Computational geometry*. 2009, pp. 237–246.

[16] Frédéric Chazal et al. *The structure and stability of persistence modules*. Springer, 2016.

[17] Tiffany J Chen and Nikesh Kotecha. "Cytobank: providing an analytics platform for community cytometry data analysis and collaboration". In: *High-dimensional single cell analysis* (2014), pp. 127–157.

[18] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. "Stability of persistence diagrams". In: *Discrete & computational geometry* 37.1 (2007), pp. 103–120.

[19] Francis H Crick. "On protein synthesis". In: *Symp Soc Exp Biol*. Vol. 12. 138-63. 1958, p. 8.

[20] Stephen C De Rosa et al. "11-color, 13-parameter flow cytometry: identification of human naive T cells by phenotype, function, and T-cell receptor diversity". In: *Nature medicine* 7.2 (2001), pp. 245–248.

[21] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[22] Emilie Devijver and Mélina Gallopin. "Block-diagonal covariance selection for high-dimensional Gaussian graphical models". In: *Journal of the American Statistical Association* 113.521 (2018), pp. 306–314.

[23] Gail Dutton. "Machine Learning Enhances Cytometry Analysis: Cytobank's cloud platform boasts FlowSOM". In: *Genetic Engineering & Biotechnology News* 39.11 (2019), pp. 14–15.

[24] Greg Finak et al. “Merging mixture components for cell population identification in flow cytometry”. In: *Advances in bioinformatics* 2009 (2009).

[25] Walter Gilbert and Allan Maxam. “The nucleotide sequence of the lac operator”. In: *Proceedings of the National Academy of Sciences* 70.12 (1973), pp. 3581–3584.

[26] Evarist Giné and Armelle Guillou. “Rates of strong uniform consistency for multivariate kernel density estimators”. In: *Annales de l’Institut Henri Poincare (B) Probability and Statistics*. Vol. 38. 6. Elsevier. 2002, pp. 907–921.

[27] Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2014.

[28] Alexander Goldenshluger and Oleg V Lepski. “Minimax estimation of norms of a probability density: I. Lower bounds”. In: *Bernoulli* 28.2 (2022), pp. 1120–1154.

[29] Alexander Goldenshluger and Oleg V Lepski. “Minimax estimation of norms of a probability density: II. Rate-optimal estimation procedures”. In: *Bernoulli* 28.2 (2022), pp. 1155–1178.

[30] Frank T Gucker Jr et al. “A photoelectronic counter for colloidal particles”. In: *Journal of the American Chemical Society* 69.10 (1947), pp. 2422–2431.

[31] Peter Hall. “On Kullback-Leibler loss and density estimation”. In: *The Annals of Statistics* (1987), pp. 1491–1519.

[32] John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[33] Rafael Hasminskii, Ildar Ibragimov, et al. “On density estimation in the view of Kolmogorov’s ideas in approximation theory”. In: *The Annals of Statistics* 18.3 (1990), pp. 999–1010.

[34] K Sparck Jones and Cornelis Joost Van Rijsbergen. “Information retrieval test collections”. In: *Journal of documentation* (1976).

[35] Jisu Kim et al. “Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3398–3407.

[36] Tania L King et al. “The use of kernel density estimation to examine associations between neighborhood destination intensity and walking and physical activity”. In: *PLoS one* 10.9 (2015), e0137402.

[37] Dmitry Kobak and George C Linderman. “Initialization is critical for preserving global data structure in both t-SNE and UMAP”. In: *Nature biotechnology* 39.2 (2021), pp. 156–157.

[38] Georges Köhler and Cesar Milstein. "Continuous cultures of fused cells secreting antibody of predefined specificity". In: *nature* 256.5517 (1975), pp. 495–497.

[39] Teuvo Kohonen. *Self-organizing maps*. Vol. 30. Springer Science & Business Media, 2012.

[40] Harold W Kuhn. "The Hungarian method for the assignment problem". In: *Naval research logistics quarterly* 2.1-2 (1955), pp. 83–97.

[41] Mark J van der Laan, Sandrine Dudoit, and Sunduz Keles. "Asymptotic optimality of likelihood-based cross-validation". In: *Statistical Applications in Genetics and Molecular Biology* 3.1 (2004).

[42] Oleg Lepski. "Multivariate density estimation under sup-norm loss: oracle approach, adaptation and independence structure". In: *Annals of Statistics* 41.2 (2013), pp. 1005–1034.

[43] Jacob H Levine et al. "Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis". In: *Cell* 162.1 (2015), pp. 184–197.

[44] Han Liu, Larry Wasserman, and John Lafferty. "Exponential concentration for mutual information estimation with application to forests". In: *Advances in Neural Information Processing Systems* 25 (2012).

[45] Han Liu et al. "Forest density estimation". In: *The Journal of Machine Learning Research* 12 (2011), pp. 907–951.

[46] Xiao Liu et al. "A comparison framework and guideline of clustering methods for mass cytometry data". In: *Genome biology* 20.1 (2019), pp. 1–18.

[47] Kenneth Lo et al. "flowClust: a Bioconductor package for automated gating of flow cytometry data". In: *BMC bioinformatics* 10.1 (2009), pp. 1–8.

[48] J MacQueen. "Classification and analysis of multivariate observations". In: *5th Berkeley Symp. Math. Statist. Probability*. 1967, pp. 281–297.

[49] Elaine R Mardis. "DNA sequencing technologies: 2006–2016". In: *Nature protocols* 12.2 (2017), pp. 213–218.

[50] Pascal Massart. "Concentration inequalities and model selection". In: (2007).

[51] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[52] Stuart Mitchell, Stuart Mitchell Consulting, and Iain Dunning. *PuLP: A Linear Programming Toolkit for Python*. 2011.

[53] Andrew Moldavan. “Photo-electric technique for the counting of microscopical cells”. In: *Science* 80.2069 (1934), pp. 188–189.

[54] David R Morrison et al. “Branch-and-bound algorithms: A survey of recent advances in searching, branching, and pruning”. In: *Discrete Optimization* 19 (2016), pp. 79–102.

[55] Iftekhar Naim et al. “SWIFT—scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, Part 1: Algorithm design”. In: *Cytometry Part A* 85.5 (2014), pp. 408–421.

[56] Evan W Newell and Yang Cheng. “Mass cytometry: blessed with the curse of dimensionality”. In: *Nature immunology* 17.8 (2016), pp. 890–895.

[57] Nick DL Owens et al. “T cell receptor signalling inspired kernel density estimation and anomaly detection”. In: *International Conference on Artificial Immune Systems*. Springer. 2009, pp. 122–135.

[58] Lily M Park, Joanne Lannigan, and Maria C Jaimes. “OMIP-069: Forty-color full spectrum flow cytometry panel for deep immunophenotyping of major cell subsets in human peripheral blood”. In: *Cytometry Part A* 97.10 (2020), pp. 1044–1051.

[59] Emanuel Parzen. “On estimation of a probability density function and mode”. In: *The annals of mathematical statistics* 33.3 (1962), pp. 1065–1076.

[60] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[61] Stephen P Perfetto, Pratip K Chattopadhyay, and Mario Roederer. “Seventeen-colour flow cytometry: unravelling the immune system”. In: *Nature Reviews Immunology* 4.8 (2004), pp. 648–655.

[62] Louis Pujol. “ISDE: Independence Structure Density Estimation”. In: *arXiv preprint arXiv:2203.09783* (2022).

[63] Saumyadipta Pyne et al. “Automated high-dimensional flow cytometric data analysis”. In: *Proceedings of the National Academy of Sciences* 106.21 (2009), pp. 8519–8524.

[64] Huijie Qiao et al. “A cautionary note on the use of hypervolume kernel density estimators in ecological niche modelling”. In: *Global Ecology and Biogeography* 26.9 (2017), pp. 1066–1070.

[65] Gilles Rebelles. “Lp adaptive estimation of an anisotropic density under independence hypothesis”. In: *Electronic journal of statistics* 9.1 (2015), pp. 106–134.

[66] Michael Reiter et al. “Clustering of cell populations in flow cytometry data using a combination of Gaussian mixtures”. In: *Pattern Recognition* 60 (2016), pp. 1029–1040.

[67] J Paul Robinson et al. “Multispectral cytometry of single bio-particles using a 32-channel detector”. In: *Advanced biomedical and clinical diagnostic systems III*. Vol. 5692. SPIE. 2005, pp. 359–365.

[68] Lucie Rodriguez et al. “Systems-level immunomonitoring from acute to recovery phase of severe COVID-19”. In: *Cell Reports Medicine* 1.5 (2020), p. 100078.

[69] M Roederer et al. “Heterogeneous calcium flux in peripheral T cell subsets revealed by five-color flow cytometry using log-ratio circuitry”. In: *Cytometry: The Journal of the International Society for Analytical Cytology* 21.2 (1995), pp. 187–196.

[70] Mario Roederer et al. “8 color, 10-parameter flow cytometry to elucidate complex leukocyte heterogeneity”. In: *Cytometry: The Journal of the International Society for Analytical Cytology* 29.4 (1997), pp. 328–339.

[71] Murray Rosenblatt. “Remarks on Some Nonparametric Estimates of a Density Function”. In: *The Annals of Mathematical Statistics* 27.3 (1956), pp. 832–837. DOI: 10.1214/aoms/1177728190. URL: https://doi.org/10.1214/aoms/1177728190.

[72] Nikolay Samusik et al. “Automated mapping of phenotype space with single-cell data”. In: *Nature methods* 13.6 (2016), pp. 493–496.

[73] Frederick Sanger, Steven Nicklen, and Alan R Coulson. “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the national academy of sciences* 74.12 (1977), pp. 5463–5467.

[74] Richard Scheuermann et al. *FlowCAP: critical assessment of flow cytometry population identification methods (65.2)*. 2011.

[75] Matthew H Spitzer and Garry P Nolan. “Mass cytometry: single cells, many features”. In: *Cell* 165.4 (2016), pp. 780–791.

[76] Istvan P Sugar and Stuart C Sealfon. “Misty Mountain clustering: application to fast unsupervised flow cytometry gating”. In: *BMC bioinformatics* 11.1 (2010), pp. 1–14.

[77] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. 1st. Springer Publishing Company, Incorporated, 2008. ISBN: 0387790519.

[78] Aad Van Der Vaart and Jon A Wellner. “A note on bounds for VC dimensions”. In: *Institute of Mathematical Statistics collections* 5 (2009), p. 103.

[79] Aad Van Der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996. ISBN: 9780387946405. URL: https://books.google.fr/books?id=seH8dMrEgggC.

[80] Sofie Van Gassen et al. "FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data". In: *Cytometry Part A* 87.7 (2015), pp. 636–645.

[81] Chris P Verschoor et al. "An introduction to automated flow cytometry gating tools and their implementation". In: *Frontiers in immunology* 6 (2015), p. 380.

[82] Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

[83] Yingfan Wang et al. "Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization." In: *J. Mach. Learn. Res.* 22.201 (2021), pp. 1–73.

[84] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Vol. 26. Springer, 2004.

[85] Lukas M Weber and Mark D Robinson. "Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data". In: *Cytometry Part A* 89.12 (2016), pp. 1084–1096.

[86] Lisa Weijler et al. "Detecting Rare Cell Populations in Flow Cytometry Data Using UMAP". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 4903–4909.

[87] Maryam Yashtini. "On the global convergence rate of the gradient descent method for functions with Hölder continuous gradients". In: *Optimization letters* 10.6 (2016), pp. 1361–1370.

[88] Habil Zare et al. "Data reduction for spectral clustering to analyze high throughput flow cytometry data". In: *BMC bioinformatics* 11.1 (2010), pp. 1–16.

Appendix A

ISDE on synthetic Gaussian data

This appendix is dedicated to the presentation of synthetic results, in the same spirit as in section 3.3 but with data drawn from centered multivariate Gaussian distributions.

Data Generating Process The Gaussian Graphical Models (GGM) theory indicates that edges of the undirected graphical model associated with a Gaussian distribution $\mathcal{N}(0, \Sigma)$ are the non-zero entries of the precision matrix Σ^{-1}. As the inverse operator preserves the block-diagonal structure, we can easily simulate data from a multivariate Gaussian with an IS.

For a positive integer s and a real number $\sigma \in (0, 1)$ we denote by Σ_σ^s the $s \times s$ matrix whose diagonal entries are 1 and non-diagonal entries are σ. Then for a list of positive integers $S = [s_1, \dots s_K]$ we define the block diagonal matrix:

$$\Sigma_\sigma^S = \left(\begin{array}{c|c|c|c} \Sigma_\sigma^{s_1} & \mathbf{0} & \dots & \mathbf{0} \\ \hline \mathbf{0} & \Sigma_\sigma^{s_2} & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & \mathbf{0} \\ \hline \mathbf{0} & \dots & \mathbf{0} & \Sigma_\sigma^{s_K} \end{array}\right) \tag{A.0.1}$$

The distribution $\mathcal{N}\left(0, \Sigma_\sigma^S\right)$ satisfies the IS condition with partition $\left(\left\{\sum_{i=1}^{j-1} s_i + 1, \ldots, \sum_{i=1}^{j} s_i\right\}\right)_{j=1,\ldots,K}$.

Evaluation Scheme If $\hat{\Sigma}$ and Σ are respectively the estimated and the true covariance, the Kullback-Leibler risk can be explicitly, see lemma A.0.1.

Lemma A.0.1

$$\mathrm{KL}\left(\mathcal{N}(0,\Sigma)\|\mathcal{N}(0,\hat{\Sigma})\right) = \sum_{v\in \mathrm{Sp}(A)} \frac{v - \log(1+v)}{2} \tag{A.0.2}$$

where $A = (\hat{\Sigma}^{-1} - \Sigma^{-1})\Sigma$.

Proof. First of all, for a covariance matrix Σ, the density f_Σ of $\mathcal{N}(0,\Sigma)$ is given by

$$\forall x \in \mathbb{R}^d f_\Sigma(x) = \frac{1}{(2\pi)^{d/2}(\det\Sigma)^{1/2}} \exp\left(-\frac{1}{2} x^{\mathrm{T}} \Sigma^{-1} x\right). \tag{A.0.3}$$

We compute the KL divergence between f_{Σ_1} and f_{Σ_2}

$$\begin{aligned}
\mathrm{KL}\left(f_{\Sigma_1}\|f_{\Sigma_2}\right) &= \int \log\left(\frac{f_{\Sigma_1}(x)}{f_{\Sigma_2}(x)}\right) f_{\Sigma_1}(x)dx && \text{(A.0.4)}\\
&= \frac{1}{2}\log\frac{\det\Sigma_2}{\det\Sigma_1} \underbrace{\int f_{\Sigma_1}(x)dx}_{=1} && \text{(A.0.5)}\\
&\quad + \frac{1}{2}\underbrace{\int x^{\mathrm{T}}\Sigma_2^{-1}x f_{\Sigma_1}(x)dx}_{=\mathrm{Tr}(\Sigma_2^{-1}\Sigma_1)} && \text{(A.0.6)}\\
&\quad + \frac{1}{2}\underbrace{\int x^{\mathrm{T}}{\Sigma_1}^{-1}x f_{\Sigma_1}(x)dx}_{=\mathrm{Tr}({\Sigma_1}^{-1}\Sigma_1)=d} && \text{(A.0.7)}\\
&= \frac{1}{2}\left(\log\det\Sigma_2 - \log\det\Sigma_1 + \mathrm{Tr}\left(\Sigma_2^{-1}\Sigma_1\right) - d\right) && \text{(A.0.8)}
\end{aligned}$$

We remark that

$$\mathrm{Tr}\left(\Sigma_2^{-1}\Sigma_1\right) - d = \mathrm{Tr}(A) = \sum_{v \in \mathrm{Sp}(A)} v \tag{A.0.9}$$

We also remark that $\log\left(\frac{\det \Sigma_1}{\det \Sigma_2}\right) = \log\left(\det \Sigma_2^{-1}\Sigma_1\right)$ and as if v is an eigenvalue of A, $1+v$ is an eigenvalue of $\Sigma_2^{-1}\Sigma_1$ we have

$$\log\left(\frac{\det \Sigma_1}{\det \Sigma_2}\right) = \sum_{v \in \mathrm{Sp}(A)} \log(1+v) \tag{A.0.10}$$

Combining these results with eq. (A.0.8) leads to the desired formula. □

Benchmarked Methods Two methods will be compared to ISDE for the task of covariance estimation.

The first estimator is the simple empirical covariance, which is the maximum likelihood estimator if the covariance does not enjoy any particular structure.

The second estimator is Block-Diagonal Covariance Selection (BDCS) developed in [22]. It aims to estimate an IS in the context of GGM. This algorithm works in two steps:

- Compute a family of nested partitions candidates to be the IS
- Choose a partition in this family using a slope heuristic approach

More details can be found in the original paper. Up to our knowledge, this is the only work dealing specifically with IS in the GGM framework.

ISDE Inputs We run algorithm 2 with $k = d$, $m = n = 0.5 \times N$ and simple empirical covariance as multivariate density estimator.

Performance We compare the three methods described above for fixed σ, N, and different structures S. We have gathered results in terms of KL loss are in table A.1. We have repeated each experiment 5 times, and the scores displayed are the mean KL losses and standard deviation over these repetitions.

S	[2, 2]	[4, 4, 1]	[4, 3, 2, 3]	[4, 4, 3, 3 ,2]
ISDE	$\mathbf{0.60 \pm 0.21}$	1.88 ± 0.52	2.85 ± 0.60	5.30 ± 0.96
BDCS	$\mathbf{0.60 \pm 0.21}$	$\mathbf{1.72 \pm 0.46}$	$\mathbf{2.63 \pm 1.01}$	$\mathbf{4.42 \pm 1.80}$
Empirical	0.80 ± 0.20	3.62 ± 0.53	6.88 ± 0.84	12.63 ± 0.83

Table A.1: Gaussian: KL Losses ($.10^3$) - $\sigma = 0.7$, $N = 6000$

Recovery We are interested not only in performance, but we also want to find the correct partition in order to get qualitative information about datasets. In table A.2 we collect, for the same experiment as above, the rate of recovery of the proper partition. In parentheses is displayed the rate of admissible output partition: a partition is admissible if all the blocks of the original partition are subsets of blocks of this one.

S	[2, 2]	[4, 4, 1]	[4, 3, 2, 3]	[4, 4, 3, 3 ,2]
ISDE	100%(100%)	80%(100%)	40%(100%)	0%(100%)
BDCS	100%(100%)	100%(100%)	80%(100%)	60%(100%)

Table A.2: Gaussian: Recovery - $\sigma = 0.7$, $N = 6000$

Conclusion We remark that BDCS is the most efficient method for the task of density estimation in GGM under IS. We can explain it as ISDE tends to select admissible partition but fails to select the exact IS when the dimension grows. BDCS inherently penalizes more useless blocks merging, making it more accurate in this setting.

However, ISDE performs significantly better than a naive empirical covariance, proving that it benefits from the IS.

We want to highlight the difference between ISDE and BDCS. BDCS starts by selecting a family of up to d nested partitions and then selects among them. This approach uses a preliminary covariance estimator to design this family of

nested partitions. This approach is reasonable as for Gaussian data, pairwise dependencies entirely determine multidimensional dependencies between features. Outside the scope of GGM, this approach does not remain valid, as features of a random variable can be pairwise independent but mutually dependent. ISDE can handle more general settings, as it selects among a set of partitions with blocks of cardinal potentially more significant than 2.

Appendix B

Bias study for the IS model in the Gaussian case

In this appendix, we study the bias $\mathrm{KL}\left(f \| f_{\mathcal{P}_*}\right)$ in a multivariate Gaussian framework. In this situation, exact computations are possible.

B.1 Model and notations

Model If Σ denotes a $d \times d$ definite positive matrix, we denote by f_Σ the density of a multivariate centered Gaussian random variable with covariance Σ and by $\Sigma_{\mathcal{P}}$ the matrix defined as follows.

$$\Sigma_{\mathcal{P}}(i,j) = \begin{cases} \Sigma(i,j) & \text{if } i \text{ and } j \text{ belongs to the same block in } \mathcal{P} \\ 0 & \text{else.} \end{cases} \tag{B.1.1}$$

If S_1 and S_2 are subsets of $\{1,\dots,d\}$, we denote by $\Sigma(S_1,S_2)$ the $|S_1| \times |S_2|$ submatrix matrix of Σ where we keep the intersection of rows in S_1 and columns in S_2, to keep notations compact, we write $\Sigma(S)$ instead of $\Sigma(S,S)$

For a multivariate Gaussian random variable with covariance Σ, $f_\Sigma \in \mathcal{D}_d^k$ is equivalent to the fact that it exists a permutation σ of $\{1, \dots, d\}$ such that $P_\sigma \Sigma P_\sigma^{-1}$ is block-diagonal with blocks of size smaller than $k \times k$. For clarity, in what follows, we will always consider that this property is met with $\sigma = \mathrm{id}$, meaning that we restrict ourselves to partitions in which each block is made of consecutive features. This does not imply a loss of generality.

We now consider that $f = f_\Sigma$ and

$$\Sigma = \Sigma_{\mathcal{P}} + \epsilon \tag{B.1.2}$$

where $\Sigma_{\mathcal{P}}$ is a block-diagonal covariance matrix corresponding to the independence structure $\mathcal{P}$ and ϵ is a "small" (in a sense to be defined later) definite positive matrix. The question is how this perturbation influences the bias term. In order to answer it, we must control $\mathrm{KL}\left(f \| f_{\mathcal{P}}\right)$ for all $\mathcal{P}$ in Part_d^k.

B.2 Some useful lemmas

Computation of KL losses The first useful result is an explicit computation of $\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_{\mathcal{P}}}\right)$ for any $\mathcal{P}$ in Part_d^k.

Lemma B.2.1

For every $\mathcal{P} \in \mathrm{Part}_d^k$

$$\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_{\mathcal{P}}}\right) = \frac{1}{2}\left(\sum_{S \in \mathcal{P}} \log\det \Sigma(S) - \log\det \Sigma\right) \tag{B.2.1}$$

Or if $\lambda_1 \leq \cdots \leq \lambda_d$ are the eigenvalues of Σ and $\lambda_1^{\mathcal{P}} \leq \cdots \leq \lambda_d^{\mathcal{P}}$ the eigenvalues of $\Sigma_{\mathcal{P}}$

$$\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_{\mathcal{P}}}\right) = \frac{1}{2}\sum_{i=1}^{d} \log\left(\frac{\lambda_i^{\mathcal{P}}}{\lambda_i}\right) \tag{B.2.2}$$

Proof. The density f_Σ has the following expression.

$$f_\Sigma(x) = \frac{1}{(2\pi)^{d/2}(\det\Sigma)^{1/2}} \exp\left(-\frac{1}{2}x^{\mathrm{T}}\Sigma^{-1}x\right). \tag{B.2.3}$$

We compute the KL divergence between f_Σ and $f_{\Sigma_\mathcal{P}}$

$$\begin{aligned}
\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_\mathcal{P}}\right) &= \int \log\left(\frac{f_\Sigma(x)}{f_{\Sigma_\mathcal{P}}(x)}\right) f_\Sigma(x)dx && \text{(B.2.4)}\\
&= \frac{1}{2}\log\frac{\det\Sigma_\mathcal{P}}{\det\Sigma}\underbrace{\int f_\Sigma(x)dx}_{=1}\\
&\quad + \frac{1}{2}\underbrace{\int x^{\mathrm{T}}\Sigma_\mathcal{P}^{-1}x f_\Sigma(x)dx}_{=\mathrm{Tr}(\Sigma_\mathcal{P}^{-1}\Sigma)}\\
&\quad + \frac{1}{2}\underbrace{\int x^{\mathrm{T}}\Sigma^{-1}x f_\Sigma(x)dx}_{=\mathrm{Tr}(\Sigma^{-1}\Sigma)=d} && \text{(B.2.5)}\\
&= \frac{1}{2}\left(\log\det\Sigma_\mathcal{P} - \log\det\Sigma + \mathrm{Tr}\left(\Sigma_\mathcal{P}^{-1}\Sigma\right) - d\right) && \text{(B.2.6)}
\end{aligned}$$

Now, we define a permutation $\sigma_\mathcal{P}$ of $\{1,\dots,d\}$ such that :

$$\Sigma_\mathcal{P} = P_{\sigma_\mathcal{P}}\left(\begin{array}{c|c|c|c} \Sigma(S_1) & \mathbf{0} & \dots & \mathbf{0} \\ \hline \mathbf{0} & \Sigma(S_2) & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & \mathbf{0} \\ \hline \mathbf{0} & \dots & \mathbf{0} & \Sigma(S_M) \end{array}\right) P_{\sigma_\mathcal{P}}^{-1} \tag{B.2.7}$$

where $\{S_1,\dots,S_M\}$ denotes the blocks of $\mathcal{P}$. It is then clear that $\log\det\Sigma_\mathcal{P} = \sum_{S\in\mathcal{P}}\log\det\Sigma(S)$. We also have

$$\Sigma = P_{\sigma_\mathcal{P}}\left(\begin{array}{c|c|c|c} \Sigma(S_1) & \Sigma(S_1,S_2) & \dots & \Sigma(S_1,S_M) \\ \hline \Sigma(S_2,S_1) & \Sigma(S_2) & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & \Sigma(S_{M-1},S_M) \\ \hline \Sigma(S_M,S_1) & \dots & \Sigma(S_M,S_{M-1}) & \Sigma(S_M) \end{array}\right) P_{\sigma_\mathcal{P}}^{-1}. \tag{B.2.8}$$

Then

$$\Sigma_{\mathcal{P}}^{-1}\Sigma = P_{\sigma\mathcal{P}} \left(\begin{array}{c|c|c|c} I_{|S_1|} & X_{1,2} & \dots & X_{1,M} \\ \hline X_{2,1} & I_{|S_2|} & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & X_{M-1,M} \\ \hline X_{M,1} & \dots & \Sigma(S_M, S_{M-1}) & I_{|S_M|} \end{array}\right) P_{\sigma\mathcal{P}}^{-1} \tag{B.2.9}$$

where for $i \neq j$, $X_{i,j}$ is a $|S_i| \times |S_j|$ matrix. Then $\mathrm{Tr}\left(\Sigma_{\mathcal{P}}^{-1}\Sigma\right) = d$.

The formulation of the result involving the eigenvalues comes from the fact that $\det \Sigma = \prod_{i=1}^{d} \lambda_i$ and $\det \Sigma_{\mathcal{P}} = \prod_{i=1}^{d} \lambda_i^{\mathcal{P}}$. □

Some computation of determinants We define the $p \times p$ matrix

$$A_\sigma^p = \begin{pmatrix} 1 & \sigma & \dots & \sigma \\ \sigma & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma \\ \sigma & \dots & \sigma & 1 \end{pmatrix}. \tag{B.2.10}$$

If k divides d we define

$$\Sigma_\sigma^{(d,k)} = \left(\begin{array}{c|c|c|c} A_\sigma^k & \mathbf{0} & \dots & \mathbf{0} \\ \hline \mathbf{0} & \ddots & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & \mathbf{0} \\ \hline \mathbf{0} & \dots & \mathbf{0} & A_\sigma^k \end{array}\right) \tag{B.2.11}$$

For $\epsilon > 0$ we define

$$\Sigma_{\sigma,\epsilon}^{(d,k)} = \left(\begin{array}{c|c|c|c} A_\sigma^k & \epsilon & \dots & \epsilon \\ \hline \epsilon & \ddots & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & \epsilon \\ \hline \epsilon & \dots & \epsilon & A_\sigma^k \end{array}\right) \tag{B.2.12}$$

Lemma B.2.2

i $\det A_\sigma^p = [1-\sigma]^{p-1}\,[1+(p-1)\sigma]$

ii $\det \Sigma_\sigma^{(d,k)} = [1-\sigma]^{\frac{d}{k}(k-1)}[1+(k-1)\sigma]^{\frac{d}{k}}$

iii $\det \Sigma_{\sigma,\epsilon}^{(d,k)} = [1-\sigma]^{\frac{d}{k}(k-1)}[1+(k-1)\sigma+(d-k)\epsilon][1+(k-1)\sigma-k\epsilon]^{\frac{d}{k}-1}$

Proof. (i) We start by computing the eigenvalues of A_σ^p. We remark that

$$A-(1-\sigma)I = \begin{pmatrix} \sigma & \dots & \sigma \\ \vdots & & \vdots \\ \sigma & \dots & \sigma \end{pmatrix}. \tag{B.2.13}$$

Then it is clear that $x \in \mathbb{R}^p \in \ker\left(A_\sigma^p-(1-\sigma)I\right) \Leftrightarrow x \in \{y \in \mathbb{R}^p : \sum_{i=1}^p y_i = 0\}$, which is a linear subspace of $\mathbb{R}^p$ of dimension $p-1$.

Then we remark that if $\mathbb{1}_p = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ then

$$A_\sigma^p \mathbb{1}_p = (1+(p-1)\sigma)\,\mathbb{1}_p \tag{B.2.14}$$

Then $1+(p-1)\sigma$ is an eigenvalue of A, and it could not be of multiplicity greater than 1 as we have just proven than $(1-\sigma)$ has a multiplicity of $p-1$. Using the fact that the determinant is the product of the eigenvalues, we obtain

$$\det A_\sigma^p = (1-\sigma)^{p-1}\,(1+(p-1)\sigma) \tag{B.2.15}$$

The proof of (ii) follows immediately, as the determinant of a block-diagonal matrix is the product of the determinants of the diagonal blocks.

(iii) We determine the eigenvalues of $\Sigma_{\sigma,\epsilon}^{(d,k)}$. To this end, we will find a set of d linearly independent eigenvectors. We remark that

$$\Sigma_{\sigma,\epsilon}^{(d,k)} \mathbb{1}_d = (1+(k-1)\sigma+(d-k)\epsilon)\mathbb{1}_d. \tag{B.2.16}$$

Then $1+(k-1)\sigma+(d-k)\epsilon$ is an eigenvalue of $\Sigma_{\sigma,\epsilon}^{(d,k)}$ with multiplicity at least 1. Now, we remark that for all integer $0 \leq i \leq \frac{d}{k}-1$ and $2 \leq j \leq k$ we have

$$\Sigma_{\sigma,\epsilon}^{(d,k)}(e_{ik+1}-e_{ik+j}) = (1-\sigma)(e_{ik+1}-e_{ik+j}). \tag{B.2.17}$$

Then $(1-\sigma)$ is an eigenvalue of $\Sigma_{\sigma,\epsilon}^{(d,k)}$ with multiplicity at least $\frac{d}{k}(k-1)$. Finally, if for $i<j$ we denote by $\mathbb{1}_j^i = \sum_{k=i}^j e_k$, we remark that for all integer $1 \leq i \leq \frac{d}{k}-1$

$$\Sigma_{\sigma,\epsilon}^{(d,k)}\left(\mathbb{1}_k^0 - \mathbb{1}_{(i+1)k}^{ik+1}\right) = (1+(k-1)\sigma-k\epsilon)\left(\mathbb{1}_k^0 - \mathbb{1}_{(i+1)k}^{ik+1}\right). \tag{B.2.18}$$

Then $1+(k-1)\sigma+(d-k)\epsilon$ is an eigenvalue of $\Sigma_{\sigma,\epsilon}^{(d,k)}$ with multiplicity at least $\frac{d}{k}-1$.

As $1+\frac{d}{k}(k-1)+\left(\frac{d}{k}-1\right)=d$, we know that the eigenvalues of $\Sigma_{\sigma,\epsilon}^{(d,k)}$ are $1+(k-1)\sigma+(d-k)\epsilon)$, $1-\sigma$ and $1+(k-1)\sigma+(d-k)\epsilon$ with multiplicity 1, $\frac{d}{k}(k-1)$ and $\frac{d}{k}-1$.

□

B.3 Control of the bias

Almost independence structure the following property precise the KL loss between $f_{\Sigma_{\sigma,\epsilon}^{(d,k)}}$ and $f_{\Sigma_\sigma^{(d,k)}}$, with a particular look at the situation where $\epsilon \to 0$.

Proposition B.3.1

$$\begin{aligned}\mathrm{KL}\left(f_{\Sigma_{\sigma,\epsilon}^{(d,k)}} \| f_{\Sigma_\sigma^{(d,k)}}\right) = &-\frac{1}{2}\log\left(1+\frac{d-k}{1+(k-1)\sigma}\epsilon\right)\\ &-\frac{1}{2}\left(\frac{d}{k}-1\right)\log\left(1-\frac{k}{1+(k-1)\sigma}\epsilon\right)\end{aligned} \tag{B.3.1}$$

At the limit $\epsilon \to 0$

$$\mathrm{KL}\left(f_{\Sigma_{\sigma,\epsilon}^{(d,k)}} \| f_{\Sigma_\sigma^{(d,k)}}\right) \underset{\epsilon\to 0}{=} \frac{d(d-k)}{4(1+(k-1)\sigma)^2}\epsilon^2 + o(\epsilon^2) \tag{B.3.2}$$

Proof. As $\Sigma_\sigma^{(d,k)}$ is a block-diagonal sub-matrix of $\Sigma_{\sigma,\epsilon}^{(d,k)}$, using lemma B.2.1 we have

$$\mathrm{KL}\left(\Sigma_{\sigma,\epsilon}^{(d,k)} \| \Sigma_\sigma^{(d,k)}\right) = \frac{1}{2}\log\left(\frac{\det \Sigma_\sigma^{(d,k)}}{\det \Sigma_{\sigma,\epsilon}^{(d,k)}}\right) \tag{B.3.3}$$

Let $\beta = 1 + (k-1)\sigma$ Now, using lemma B.2.2, we have

$$\mathrm{KL}\left(\Sigma_{\sigma,\epsilon}^{(d,k)} \| \Sigma_\sigma^{(d,k)}\right) = \frac{1}{2}\left[\log\left(\frac{\beta^{\frac{d}{k}}}{[\beta + (d-k)\epsilon][\beta - k\epsilon]^{\frac{d}{k}-1}}\right)\right] \tag{B.3.4}$$

$$= \frac{1}{2}\left[\frac{d}{k}\log\beta - \log(\beta + (d-k)\epsilon) - \left(\frac{d}{k} - 1\right)\log\left(\beta - k\epsilon\right)\right] \tag{B.3.5}$$

$$= \frac{d}{2k}\log\beta - \frac{\log\beta}{2} - \frac{1}{2}\log\left(1 + \frac{d-k}{\beta}\epsilon\right)$$

$$- \frac{1}{2}\left(\frac{d}{k} - 1\right)\log\beta - \frac{1}{2}\left(\frac{d}{k} - 1\right)\log\left(1 - \frac{k}{\beta}\epsilon\right) \tag{B.3.6}$$

$$= -\frac{1}{2}\log\left(1 + \frac{d-k}{\beta}\epsilon\right) - \frac{1}{2}\left(\frac{d}{k} - 1\right)\log\left(1 - \frac{k}{\beta}\epsilon\right) \tag{B.3.7}$$

Now, we use that

$$\log(1 + \frac{d-k}{\beta}\epsilon) \underset{\epsilon\to 0}{=} \frac{d-k}{\beta}\epsilon - \frac{(d-k)^2}{2\beta^2}\epsilon^2 + o(\epsilon^2) \tag{B.3.8}$$

and

$$\log(1 - \frac{k}{\beta}\epsilon) \underset{\epsilon\to 0}{=} -\frac{k}{\beta}\epsilon - \frac{k^2}{2\beta^2}\epsilon^2 + o(\epsilon^2). \tag{B.3.9}$$

And we obtain

$$\mathrm{KL}\left(\Sigma_{\sigma,\epsilon}^{(d,k)} \| \Sigma_\sigma^{(d,k)}\right) \underset{\epsilon\to 0}{=} -\frac{d-k}{2\beta}\epsilon + \frac{(d-k)^2}{4\beta^2}\epsilon^2$$

$$+ \frac{\left(\frac{d}{k} - 1\right)k}{2\beta}\epsilon + \frac{\left(\frac{d}{k} - 1\right)k^2}{4\beta^2}\epsilon^2 + o(\epsilon^2) \tag{B.3.10}$$

$$\underset{\epsilon\to 0}{=} \frac{(d-k)^2 + kd - k^2}{4\beta^2}\epsilon^2 + o(\epsilon^2) \tag{B.3.11}$$

$$\underset{\epsilon\to 0}{=} \frac{d^2 - 2kd + k^2 + kd - k^2}{4\beta^2}\epsilon^2 + o(\epsilon^2) \tag{B.3.12}$$

$$\underset{\epsilon\to 0}{=} \frac{d(d-k)}{4\beta^2}\epsilon^2 + o(\epsilon^2) \tag{B.3.13}$$

□

Optimal structure for small k The following proposition establishes that if $\Sigma = A_\sigma^d$ and $k < d$, $\mathcal{P}_*$ is composed of a maximum number of blocks of size k.

Proposition B.3.2

Suppose that $\Sigma = A_\sigma^d$. A structure $s = (s_i)_{i=1}^M$ is a list of positive integer with $\sum_{i=1}^M s_i = d$. To a structure is associated a partition with blocks of consecutive features with size $s_1, \dots, s_M$. For any structure s we have

$$\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_s}\right) = \frac{1}{2}\left(\sum_{i=1}^{M} \log\left(\frac{1+(s_i-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(d-1)\sigma}{1-\sigma}\right)\right). \tag{B.3.14}$$

If we denote by p and r the only integers such that $d = pk + r$ where $r < k$, we have

$$s^* = (\underbrace{k, \dots, k}_{p \text{ times}}, r) \tag{B.3.15}$$

Proof. We combine lemma B.2.1 and lemma B.2.2 to obtain

$$\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_s}\right) = \frac{1}{2}\left(\sum_{S\in\mathcal{P}} \log\det A_\sigma^{s_i} - \log\det A_\sigma^d\right) \tag{B.3.16}$$

$$= \frac{1}{2}[\sum_{S\in\mathcal{P}}(s_i-1)\log(1-\sigma) + \log(1+(s_i-1)\sigma) - (d-1)\log(1-\sigma) - \log(1+(d-1)\sigma)] \tag{B.3.17}$$

$$= \frac{1}{2}\left(\underbrace{\sum_{i=1}^{M}(s_i-1)}_{=d-M} - (d-1)\right)\log(1-\sigma) + \frac{1}{2}\left(\sum_{i=1}^{M}\log\left(1+(s_i-1)\sigma\right) - \log(1+(d-1)\sigma)\right) \tag{B.3.18}$$

$$= \frac{1}{2}\left(\sum_{i=1}^{M}\log\left(\frac{1+(s_i-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(d-1)\sigma}{1-\sigma}\right)\right) \tag{B.3.19}$$

Now we want to prove that the structure minimizing $\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_s}\right)$ is $(\underbrace{k,\ldots,k}_{p \text{ times}}, r)$. To do so we start by remarking that for any $s = (s_i)_{i=1}^M \neq (k,\ldots,k,r)$ it exists $i \neq j$ such that $s_i \neq k$ and $s_j \neq k$. We will prove that, for our minimization problem, it is always possible to find a better structure $\tilde{s}$ with the following

i if $s_i + s_j \leq k$, $\tilde{S} = (s_k)_{k\notin\{i,j\}} \cup (s_i + s_j)$

ii if $\exists l > 0 : s_i + s_j = k + l$, $\tilde{S} = (s_k)_{k\notin\{i,j\}} \cup (k, l)$

To prove (i), we start from

$$2\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_s}\right) = \sum_{k=1}^{M}\log\left(\frac{1+(s_k-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(d-1)\sigma}{1-\sigma}\right) \tag{B.3.20}$$

$$2\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_{\tilde{s}}}\right) = \sum_{k=1,\ldots,M,k\notin\{i,j\}} \log\left(\frac{1+(s_k-1)\sigma}{1-\sigma}\right) + \log\left(\frac{1+(s_i+s_j-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(d-1)\sigma}{1-\sigma}\right). \tag{B.3.21}$$

Then to prove that $\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_s}\right) > \mathrm{KL}\left(f_\Sigma \| f_{\Sigma_{\tilde{s}}}\right)$ it is sufficient to prove that for all $a, b \geq 1$, $g(a,b) > 0$ where

$$g(a,b) = \log\left(\frac{1+(a-1)\sigma}{1-\sigma}\right) + \log\left(\frac{1+(b-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(a+b-1)\sigma}{1-\sigma}\right). \tag{B.3.22}$$

Let us start by computing $\partial_1 g(a,b)$

$$\partial_1 g(a,b) = \frac{\sigma}{1+(a-1)\sigma} - \frac{\sigma}{1+(a+b-1)\sigma} \tag{B.3.23}$$

$$= \frac{\sigma}{(1+(a-1)\sigma)\,(1+(a+b-1)\sigma)}(1+(a+b-1)\sigma - 1 - (a-1)\sigma) \tag{B.3.24}$$

$$= \frac{b\sigma^2}{(1+(a-1)\sigma)\,(1+(a+b-1)\sigma)} \geq 0. \tag{B.3.25}$$

Then for any $b \geq 1$, $g(a,b)$ is non-decreasing in a. As a and b play similar roles in $g(a,b)$, we have that for any $a \geq 1$, $g(a,b)$ is non-decreasing in b. To prove that $g(a,b) \geq 0$ it is sufficient to show that $g(1,1) > 0$.

$$g(1,1) = \log\left(\frac{1}{1-\sigma}\right) + \log\left(\frac{1}{1-\sigma}\right) - \log\left(\frac{1+\sigma}{1-\sigma}\right) \tag{B.3.26}$$

$$= -\log\left((1-\sigma)(1+\sigma)\right). \tag{B.3.27}$$

Now, as $\sigma \in (0,1)$, $(1-\sigma)(1+\sigma) \in (0,1)$. Then $\log\left((1-\sigma)(1+\sigma)\right) < 0$, implying $g(1,1) > 0$.

To prove (ii) we start from

$$2\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_s}\right) = \sum_{k=1}^{M} \log\left(\frac{1+(s_k-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(d-1)\sigma}{1-\sigma}\right) \tag{B.3.28}$$

$$2\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_{\tilde{s}}}\right) = \sum_{k=1,\ldots,M,k\notin\{i,j\}} \log\left(\frac{1+(s_k-1)\sigma}{1-\sigma}\right) + \log\left(\frac{1+(k-1)\sigma}{1-\sigma}\right) + \log\left(\frac{1+(l-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(d-1)\sigma}{1-\sigma}\right). \tag{B.3.29}$$

Then to prove that $\mathrm{KL}\left(f_\Sigma \| f_{\Sigma_s}\right) > \mathrm{KL}\left(f_\Sigma \| f_{\Sigma_{\tilde{s}}}\right)$ it is sufficient to prove that for all $x \in [l,k]$, $h(x)$ attains its minimum at l or k where

$$h(x) = \log\left(1+(x-1)\sigma\right) + \log\left(1+((k+l)-x-1)\sigma\right). \tag{B.3.30}$$

Let us start by computing $h'(x)$

$$h'(x) = \sigma\left(\frac{1}{1+(x-1)\sigma} - \frac{1}{1+((k+l)-x-1)\sigma}\right) \tag{B.3.31}$$

$$= \frac{\sigma\left(1+((k+l)-x-1)\sigma - 1 - (x-1)\sigma\right)}{(1+(x-1)\sigma)(1+((k+l)-x-1)\sigma)} \tag{B.3.32}$$

$$= \frac{\sigma^2}{(1+(x-1)\sigma)(1+((k+l)-x-1)\sigma)}((k+l)-2x). \tag{B.3.33}$$

Then h increases from l to $(k+l)/2$ and decreases from $(k+l)/2$ to k and $h(l) = h(k)$ the minimum of h is attained on l and k.

□

Conclusion We finish this appendix by establishing a general upper bound of $\mathrm{KL}\left(f_\Sigma \| f_{\mathcal{P}_*}\right)$ where $\Sigma = \Sigma_{\sigma,\epsilon}^{(d,k^*)}$.

Theorem B.3.3

If $\Sigma = \Sigma_{\sigma,\epsilon}^{(d,k^*)}$ and if $k < k^*$. Let (p, r) be the unique couple of integer with $0 \leq r < k$ such that $k^* = pk + r$, we have

$$\mathrm{KL}\left(f_\Sigma \| f_{\mathcal{P}_*}\right) \leq \mathrm{KL}\left(f_{\Sigma_{\sigma,\epsilon}^{(d,k^*)}} \| f_{\Sigma_\sigma^{(d,k^*)}}\right) + \frac{dp}{2k^*}\log\left(\frac{1+(k-1)\sigma}{1-\sigma}\right) + \frac{d}{2k^*}\log\left(\frac{1+(r-1)\sigma}{1-\sigma}\right) - \frac{d}{2k^*}\log\left(\frac{1+(k^*-1)\sigma}{1-\sigma}\right). \tag{B.3.34}$$

If $k \geq k^*$, we have

$$\mathrm{KL}\left(f_\Sigma \| f_{\mathcal{P}_*}\right) \leq \mathrm{KL}\left(f_{\Sigma_{\sigma,\epsilon}^{(d,k^*)}} \| f_{\Sigma_\sigma^{(d,k^*)}}\right). \tag{B.3.35}$$

And

$$\mathrm{KL}\left(f_{\Sigma_{\sigma,\epsilon}^{(d,k)}} \| f_{\Sigma_\sigma^{(d,k)}}\right) \underset{\epsilon\to 0}{=} \frac{d(d-k)}{4(1+(k-1)\sigma)^2}\epsilon^2 + o(\epsilon^2) \tag{B.3.36}$$

Proof. If $k < k^*$, let us consider

- the structure $\tilde{s} = (\underbrace{k, \dots, k}_{p \text{ times}}, r)$, and $\mathcal{P}_{\tilde{s}}$ the associated partition of k^* features,
- the structure $s = (\underbrace{\tilde{s}, \dots, \tilde{s}}_{d/k^* \text{ times}})$, and $\mathcal{P}_s$ the associated partition of d features,
- the structure $s_0 = (\underbrace{k^*, \dots, k^*}_{d/k^* \text{ times}})$ and $\mathcal{P}_0$ the associated partition of d features.

We can upper-bound the bias term $\mathrm{KL}\left(f_\Sigma \| f_{\mathcal{P}_*}\right)$ as follows

$$\begin{aligned}
\mathrm{KL}\left(f_\Sigma \| f_{\mathcal{P}_*}\right) &\leq \mathrm{KL}\left(f_\Sigma \| f_{\mathcal{P}_s}\right) && \text{(B.3.37)}\\
&= \int \log\left(\frac{f_\Sigma}{f_{\mathcal{P}_s}}\right) f_\Sigma && \text{(B.3.38)}\\
&= \int \log\left(\frac{f_\Sigma}{f_{\mathcal{P}_0}}\right) f_\Sigma + \int \log\left(\frac{f_{\mathcal{P}_0}}{f_{\mathcal{P}_s}}\right) f_\Sigma && \text{(B.3.39)}\\
&= \mathrm{KL}\left(\Sigma_{\sigma,\epsilon}^{(d,k^*)} \| \Sigma_\sigma^{(d,k^*)}\right) + \int \log\left(\frac{f_{\mathcal{P}_0}}{f_{\mathcal{P}_s}}\right) f_\Sigma && \text{(B.3.40)}
\end{aligned}$$

The blocks of the partition $\mathcal{P}_s$ are subsets of blocks of the partition $\mathcal{P}_0$, then

$$\begin{aligned}
\int \log\left(\frac{f_{\mathcal{P}_0}}{f_{\mathcal{P}_s}}\right) f_\Sigma &= \int \log\left(\frac{\prod_{S\in\mathcal{P}_0} f_S}{\prod_{S\in\mathcal{P}_0} (f_{\mathcal{P}_s})_S}\right) f_\Sigma && \text{(B.3.41)}\\
&= \sum_{S\in\mathcal{P}_0} \int \log\left(\frac{f_S}{(f_{\mathcal{P}_s})_S}\right) f_S && \text{(B.3.42)}\\
&= \sum_{S\in\mathcal{P}_0} \mathrm{KL}\left(f_{A_\sigma^d} \| f_{(A_\sigma^d)_{\mathcal{P}_{\tilde{s}}}}\right) && \text{(B.3.43)}\\
&= \frac{d}{k^*} \mathrm{KL}\left(f_{A_\sigma^d} \| f_{(A_\sigma^d)_{\mathcal{P}_{\tilde{s}}}}\right) && \text{(B.3.44)}
\end{aligned}$$

If $k \geq k^*$, it exists $\mathcal{P} \in \mathrm{Part}_d^k$ such that $f_{\mathcal{P}} = f_{\Sigma_\sigma^{(d,k)}}$, $\mathcal{P}$ being composed of blocks of size k composed by consecutive variables. Then

$$\mathrm{KL}\left(f_\Sigma \| f_{\mathcal{P}_*}\right) \leq \ \mathrm{KL}\left(f_{\Sigma_{\sigma,\epsilon}^{(d,k^*)}} \| f_{\Sigma_\sigma^{(d,k^*)}}\right). \qquad \text{(B.3.45)}$$

Now, using proposition B.3.1

$$\mathrm{KL}\left(f_{A_\sigma^d} \| f_{(A_\sigma^d)_{\mathcal{P}_{\tilde{s}}}}\right) = \frac{1}{2}\left(p \log\left(\frac{1+(k-1)\sigma}{1-\sigma}\right) + \log\left(\frac{1+(r-1)\sigma}{1-\sigma}\right) - \log\left(\frac{1+(k^*-1)\sigma}{1-\sigma}\right)\right) \tag{B.3.46}$$

And, using proposition B.3.2, we have that

$$\mathrm{KL}\left(f_{\Sigma_{\sigma,\epsilon}^{(d,k)}} \| f_{\Sigma_\sigma^{(d,k)}}\right) \underset{\epsilon \to 0}{=} \frac{d(d-k)}{4(1+(k-1)\sigma)^2}\epsilon^2 + o(\epsilon^2). \tag{B.3.47}$$

Then we have proven the desired upper-bound.

□

The bound for the bias $\mathrm{KL}\left(f_\Sigma \| f_{\mathcal{P}_*}\right)$ where $\Sigma = \Sigma_{\sigma,\epsilon}^{(d,k^*)}$ decomposes as the sum of two terms:

- $\mathrm{KL}\left(f_{\Sigma_{\sigma,\epsilon}^{(d,k)}} \| f_{\Sigma_\sigma^{(d,k)}}\right)$ is linked to the fact that the true density does not enjoy exactly an IS but rather an approximate version were proper independence is replaced by small correlation coefficients ϵ. As showed by proposition B.3.1, these quantities goes to 0 at the rate ϵ^2 for fixed values of d, k and σ.
- The remaining part of the upper bound in the case $k < k^*$ is related to the incompressible error made by selecting a too small parameter for the size of the blocks.