# ANALYSE D'IMAGES TERAHERTZ

par

Mohamed Walid Ayech

Thèse présentée au Département d'informatique
en vue de l'obtention du grade de philosophiæ doctor (Ph.D.)

FACULTÉ DES SCIENCES

UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, Mai 2019

Le 7 mai 2019

*le jury a accepté la thèse de Monsieur Mohamed Walid Ayech dans sa version finale.*

Membres du jury

Professeur Djemel Ziou
Directeur de recherche
Département d'informatique

Professeur Taoufik Bouezmarni
Membre interne
Département de Mathématique

Professeur Wassim Bouachir
Membre externe
Département Science et Technologie
Université TÉLUQ

Professeur Richard Egli
Président-rapporteur
Département d'informatique

# Sommaire

Cette thèse à publications présente toutes nos contributions qui se rapportent à la segmentation d'images Térahertz. La thèse comprend quatre chapitres. Les deux premiers chapitres introduisent deux nouvelles approches de segmentation basées sur des techniques d'échantillonnage. Dans la première approche, nous formulons la technique de classification $K$-means dans le cadre de l'échantillon d'ensembles ordonnés pour surmonter le problème d'initialisation des centres. Le deuxième chapitre aborde la sélection des données à travers la pondération de caractéristiques et l'échantillonnage aléatoire simple pour la classification des pixels en vue d'une segmentation des images Térahertz. Une estimation automatique de la taille de l'échantillon aléatoire et du nombre de caractéristiques sélectionnées sont également proposés. Les deux chapitres suivants introduisent une autre famille de techniques de classification des séries chronologiques basées sur la régression et qui sont adaptées aux séries chronologiques. Nous supposons que les valeurs associées à chaque pixel d'une image Térahertz sont échantillonnées à partir d'un modèle autorégressif. La segmentation de l'image est alors vue comme un problème de classification de séries chronologiques. Ainsi, dans le troisième chapitre, la classification est formulée comme un problème d'optimisation non-linéaire. L'ordre du modèle et le nombre de classes sont estimés en utilisant un critère généralisé d'information. Finalement, le quatrième chapitre est une généralisation des résultats obtenus dans le troisième chapitre. Au lieu de considérer un problème de moindres carrés, nous proposons une approche de classification probabiliste basée sur le mélange de modèles autorégressifs. Les paramètres de l'approche proposée sont automatiquement estimés en utilisant un critère généralisé d'information.

# Remerciements

Durant mon doctorat, j'ai eu la chance de rencontrer plusieurs personnes. Il est arrivé le moment de les remercier de m'avoir aider lors de ce parcours. Je tiens d'abord à exprimer ma reconnaissance à mon directeur de thèse Djemel Ziou qui m'a accueilli dans son groupe, qui m'a encadré durant ces années et qui m'a donné les conseils et les moyens pour mener ce travail à bien. Son encadrement et sa disponibilité furent des facteurs importants pour la bonne conduite de mon sujet de recherche. Je voudrais également remercier les membres du jury Richard Egli, Taoufik Bouezmarni et Wassim Bouachir qui ont accepté d'évaluer ce travail.

Je suis ravi des bons moments passés avec mes collègues du laboratoire MOIVRE (MOdelisationen Imagerie, Vision et REseaux de neurones) pour leur aide précieuse. Je leur souhaite toute la réussite pour leur travail et leur recherche. J'adresse un remerciement particulier pour le professeur Béchir Ayeb, qui m'a soutenu et encouragé durant mes études.

Merci à mes parents Abdel Karim et Nabiha d'avoir cru en moi. Merci à mes frères Marouane et Zied, mes soeurs Haifa, Syrine et Safa et mes beaux-frères Ayhem et Amine pour leur appui et leur encouragement tout au long de mon doctorat. Je souhaite remercier tout spécialement Hanen qui m'a soutenu moralement dans les moments les plus difficiles. Je me sens très privilégié d'avoir une femme si exceptionnelle dans ma vie.

# Abréviations

**SRS** Simple random sampling

**RSS** Ranked set sampling

**SRS-$K$-Means** $K$-Means based SRS sampling

**Ranked-$K$-Means** $K$-Means based RSS sampling

**W-$K$-Means** $K$-Means based feature weighting

**SS-$K$-Means** $K$-Means based feature selection and random pixel sampling

**GMM** Gaussian mixture model

**KHM** K-Harmonic means

**AR** Autoregressive model

**$K$-AR** $K$-Autoregressive models

**MoAR** Mixture of Autoregressive models

# Table des matières

# Table des figures

# Liste des tableaux

## Liste des tableaux

# Introduction

Le présent chapitre introduit la technologie d'imagerie avec les radiations Térahertz. Il présente ensuite le processus de formation de cette technologie d'images et énumère les motivations pour analyser les images Térahertz. Ce chapitre décrit la feuille de route pour notre thèse et un bref résumé des contributions présentées dans notre travail.

## 1 Radiation et imagerie Térahertz

Au cours des dernières années, des nombreux groupes de recherche à travers le monde sont intéressés à la portion Térahertz (THz) des rayonnements électromagnétiques [33, 42, 63, 112]. Les rayonnements Térahertz (rayons T) se réfèrent à la région du rayonnement électromagnétique occupant la bande de fréquences de 0.1 à 10 THz (qui correspond à des longueurs d'ondes comprises entre 3 mm et 3 $\mu$m), délimitées par les micro-ondes et les ondes infrarouges (voir la figure 1). Par rapport aux régions optiques, infrarouges et rayons X, les développements technologiques avancés dans la région Térahertz sont limités. Cependant, les progrès récents dans les technologies électro-optiques rendent actuellement la région Térahertz disponible pour un usage pratique. Ces progrès technologiques ont rendu possible la génération et la détection des rayons T avec des dispositifs efficaces. Une portion inutilisée du spectre électromagnétique devient disponible, c'est une portion d'un grand potentiel pour la détection et l'imagerie en microélectronique [6, 5, 63], en diagnostic médical [54, 55, 85, 112], en contrôle environnemental [33, 83, 82], en sécurité [101, 71, 76], en identification chimique et biologique [113, 52, 102], en contrôle de qualité [14, 82, 58, 59], etc.

Depuis 1995, les rayonnements Térahertz ont offert des nouvelles possibilités pour

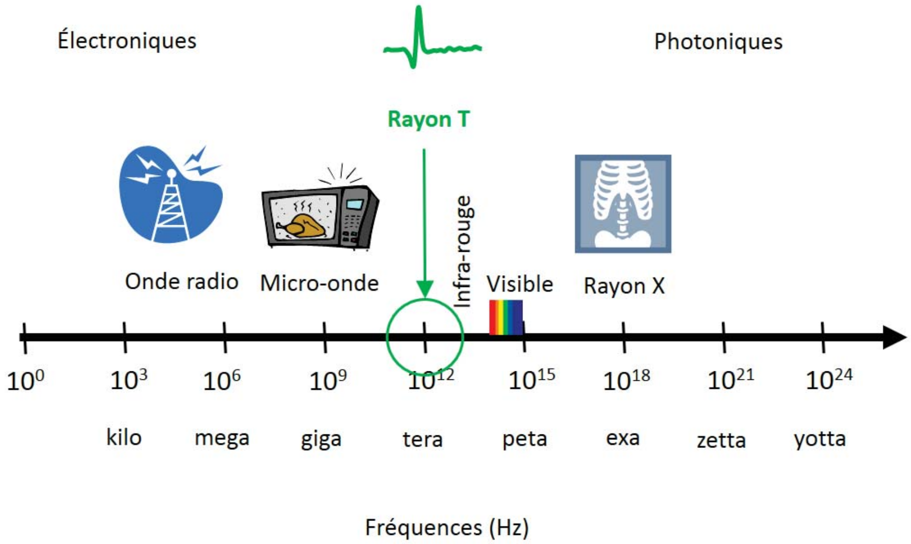


figure 1 – Spectre des ondes électromagnétiques. Les ondes Térahertz sont définies entre les micro-ondes et les ondes infrarouge.

des applications scientifiques et industrielles [42, 54, 82, 83, 99, 90, 66, 60]. Bien que d'autres portions électromagnétiques sont assez utilisées dans des applications de l'imagerie numériques, les propriétés de l'imagerie avec les rayonnements Térahertz leurs permettent d'occuper une position importante. Le tableau 1 résume les principaux avantages et inconvénients de l'imagerie avec les rayons T par rapport aux autres technologies, telles que les micro-ondes, les infrarouges et les rayons X. Les micro-ondes offrent une bonne profondeur de pénétration à travers des objets opaques, mais ils permettent une faible résolution spatiale. Les rayonnements infrarouges offrent une bonne résolution spatiale, mais ils permettent une faible profondeur de pénétration à travers les objets opaques. Les rayons X offrent la meilleure résolution spatiale et la profondeur de pénétration la plus élevée, mais ils sont potentiellement invasifs pour les organismes vivants ou les tissus biologiques inspectés. L'imagerie avec les rayons T offre un bon compromis entre les modalités mentionnées ci-dessus. Les rayons T sont moins énergétiques que les rayons X et ne semblent pas présenter aucun risque de santé pour l'inspection des organismes vivants, des tissus biologiques, des nourritures et textiles industriels [23]. Les rayons Térahertz produisent des informations spectrales non disponibles à travers d'autres types de rayons et rendent possible

| Radiations | Avantages | Inconvénients |
|---|---|---|
| Micro-ondes | - Pas dangereuses<br>- Bonne pénétration dans des nombreux matériaux<br>- Système d'acquisition d'images rapide | - Faible résolution spatiale<br>- Les métaux et l'eau bloquent le rayonnement<br>- Imagerie spectrale non disponible<br>- Coût élevé de maintenance |
| Infra-rouges | - Pas dangereuses<br>- Bonne résolution spatiale<br>- Système d'acquisition d'images rapide<br>- Coût acceptable de maintenance | - Faible profondeur de pénétration<br>- Imagerie spectrale non disponible |
| Rayons X | - Haute profondeur de pénétration<br>- Excellente résolution spatiale<br>- Système d'acquisition d'images rapide | - Dangereuses pour les êtres vivants<br>- Coût élevé de maintenance<br>- Imagerie spectrale non disponible |
| Rayons T | - Pas dangereuses<br>- Bonne profondeur de pénétration<br>- Bonne résolution spatiale<br>- Imagerie spectrale disponible | - Système d'acquisition d'images lent<br>- Les métaux et l'eau bloquent le rayonnement<br>- Coût élevé de maintenance |

tableau 1 – Résumé des principaux avantages et inconvénients de l'imagerie avec les radiations (rayons) Térahertz par rapport aux autres technologies. Tableau extrait de [95]

la discrimination des matériaux spécifiques à l'intérieur d'un objet. Par rapport aux micro-ondes, les courtes longueurs d'onde de la portion Térahertz permettent une plus grande résolution spatiale. En résumé, les rayons Térahertz sont caractérisés par plusieurs propriétés importantes, parmi lesquelles, l'inspection non invasive des objets, la pénétration à travers des objets secs et non métalliques tels que le plastique, le carton, le bois et le tissu et offrent une identification spécifique des matériaux [54, 19, 23].

Cependant, par rapport aux modalités d'imagerie bien développées, comme les infrarouges et les rayons X, les systèmes d'acquisition d'images THz ont des dispositifs d'acquisition lents. Cette lenteur s'explique par l'immaturité de cette nouvelle modalité. Dans ce qui suit, nous présentons le processus de formation d'images Térahertz, suivi par d'autres défis rencontrés par cette technologie d'images et les motivations pour son analyse et son interprétation.

# 2 Formation d'images Térahertz

L'imagerie avec les rayons Térahertz peut être obtenue par une acquisition en deux modes passif ou actif. Un système passif utilise la lumière du soleil comme une source et détecte les rayons Térahertz émis naturellement par un objet. Le système actif diffère du système passif par l'utilisation d'une source active des rayons THz artificiels pour éclairer l'objet et détecter les rayons THz transmis ou réfléchis. Dans notre travail, les acquisitions sont utilisées uniquement en mode actif.

Il y a presque quarante ans, la génération des radiations électromagnétiques est apparue en utilisant le laser à impulsions ultra-brèves [11]. Les impulsions laser femtoseconde sont utilisées pour générer des impulsions électriques picosecondes dont les bandes spectrales se trouvent dans la région Térahertz. La spectroscopie Térahertz dans le domaine temporel est d'abord utilisée pour détecter ces impulsions à large bande afin d'analyser la réponse spectrale des matériaux [114]. Cette technique est ensuite étendue à l'imagerie Térahertz, qui peut être mise en oeuvre en mode de réflexion ou de transmission. Dans la suite de ce chapitre, nous décrivons un système d'acquisition de l'image THz en mode transmission (voir figure 2). Le système d'acquisition enregistre les réponses spectroscopiques d'un échantillon qui est cartographié à plusieurs positions contigus de pixels [63]. Le système d'acquisition d'images Térahertz commence par émettre des impulsions laser ultra-rapide (typiquement entre 10 et 100 fs) vers un séparateur de faisceaux. Les impulsions laser sont divisées en faisceaux de pompe et de sonde. Le faisceau de pompe est utilisé pour générer des rayons Térahertz et le faisceau de sonde est utilisé pour détecter le champ électrique des rayons THz d'une manière cohérente. Les rayons Térahertz à large bande sont générés par l'illumination avec le faisceau de pompe dans un cristal (tel que ZnTe, GaAs et InP) avec rectification optique [54]. Des miroirs paraboliques sont nécessaires pour focaliser les rayons Térahertz produits vers un endroit où l'échantillon est situé. Les rayons Térahertz interagissent avec l'échantillon avant d'être transmis au détecteur au moyen de miroirs paraboliques. À un instant donné, le détecteur est déclenché par le faisceau de sonde et l'amplitude du champ électrique Térahertz est mesurée. L'instant de la mesure est déterminé par le retard du faisceau de sonde. Le balayage de toute l'impulsion Térahertz par la ligne à retard du faisceau de sonde permet de reconstruire

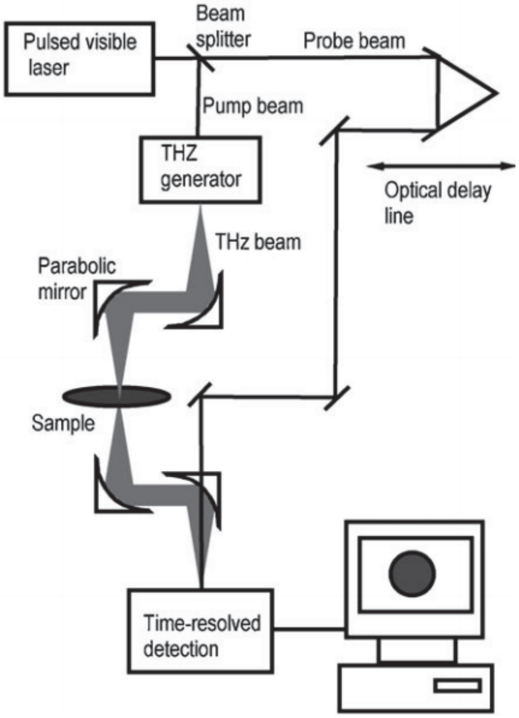
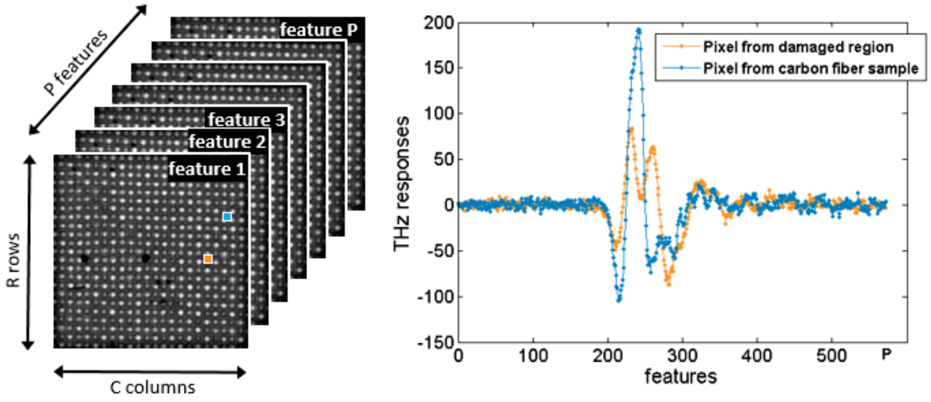


figure 2 – Schéma typique d'un système de formation d'images Térahertz en mode transmission (figure extraite de [54]). Les signaux THz sont projetés sur l'objet, interagis avec celui-ci, puis détectés pour constituer un cube de données THz. Les signaux projetés sont similaires, tandis que les signaux détectés sont modifiés qui illustrent les différentes régions de l'objet.

cette impulsion en une série de points discrets [54]. L'ensemble d'impulsions détectées sont alors enregistrées à plusieurs emplacements contigus qui constituent les pixels de l'image THz. Chaque pixel est considéré comme une série chronologique représentée par plusieurs bandes, caractéristiques ou attributs (par exemple, 1500 bandes). Ainsi, la combinaison de ces séries en lignes et en colonnes constitue un cube de données Térahertz brutes (par exemple, le cube $R \times C \times P$ dans la figure 3, où $R$, $C$ et $P$ représentent respectivement le nombre de lignes, de colonnes et de bandes). Pour visualiser l'image THz, les caractéristiques peuvent être extraites pour créer l'image 2D. On peut sélectionner l'amplitude pour un délai de temps spécifique, l'amplitude



(a) Cube 3D de données THz (b) Deux réponses (signaux) Térahertz

figure 3 – (a) Cube 3D de données Térahertz représenté par $R \times C$ pixels et caractérisé par des $P$ attributs brutes. Deux pixels colorés en bleu et en orangé appartiennent respectivement à une région typique de la fibre de carbone et à une région endommagée. (b) contient deux réponses THz différentes colorées en bleu et en orangé qui correspondent respectivement aux deux pixels de l'image Térahertz en (a) pour la même couleur.

maximale ou minimale de chaque série ou l'amplitude de la transformée de Fourier prise sur un intervalle de temps [54, 22].

# 3 Problématique de la thèse

L'imagerie dans le domaine Térahertz peut fournir des informations temporelles et spectrales spécifiques et non disponibles pour d'autres modalités. Cependant, l'imagerie Térahertz fait face à des défis pour pouvoir l'analyser et l'interpréter automatiquement. La quantité énorme de caractéristiques brutes peut être un obstacle pour décrire avec une certaine précision le contenu informationnel des images THz [112, 85, 51]. De plus, certaines caractéristiques de l'image THz brute peuvent être bruitées, redondantes ou non informatives [14]. Le nombre élevé de pixels peut aussi être une barrière pour analyser ce type d'images [61, 15]. Le traitement de l'ensemble complet des mesures nécessite une consommation élevée de la mémoire et de calcul.

L'objectif de cette thèse est de segmenter les images Térahertz en utilisant des méthodes d'analyse de données. La segmentation de ces images consiste à partitionner l'ensemble de pixels en plusieurs régions homogènes pour localiser des objets dans les images. Ces objets sont supposés disjoints et les régions qui les constituent, forment des classes séparées dans l'espace de caractéristiques. Vue la quantité énorme de caractéristiques, nous proposons dans cette thèse des stratégies de réduction de l'espace de caractéristiques. L'extraction de caractéristiques et la reconnaissance d'objets sont effectuées dans l'espace réduit. Pour ce faire, nous privilégierons d'utiliser des techniques de classification pour analyser ce type d'images. Dans la section suivante, nous présentons nos contributions relatives à l'analyse d'images Térahertz.

## 4 Contributions

Les deux premiers chapitres introduisent deux nouvelles approches de segmentation basées sur des techniques d'échantillonnage. Dans la première approche, nous formulons la technique de classification $K$-means dans le cadre de l'échantillon d'ensembles ordonnés pour surmonter le problème d'initialisation des centres. Le deuxième chapitre aborde la sélection des données à travers la pondération de caractéristiques et l'échantillonnage aléatoire simple pour la classification des pixels en vue d'une segmentation des images Térahertz. Une estimation automatique de la taille de l'échantillon aléatoire et du nombre de caractéristiques sélectionnées sont également proposés. Dans ces deux chapitres, nous avons réalisé des tests sur des ensembles de données de synthèse et d'images Térahertz qui ont permis d'évaluer la performance des méthodes proposées par rapport à l'état de l'art.

Les deux derniers chapitres introduisent une autre famille de techniques de classification des pixels basées sur la régression et qui sont adaptées aux séries chronologiques. Nous supposons que les valeurs associées à chaque pixel d'une image Térahertz sont échantillonnées à partir d'un modèle autorégressif. La segmentation de l'image est alors vue comme un problème de classification de séries chronologiques. Ainsi, dans le troisième chapitre, la classification est formulée comme un problème d'optimisation non-linéaire. L'ordre du modèle et le nombre de classes sont estimés en utilisant un critère généralisé d'information. Finalement, le quatrième chapitre est une généralisation des résultats obtenus dans le troisième chapitre. Au lieu de considérer un problème de moindres carrés, nous proposons une approche de classification probabiliste basée sur le mélange de modèles autorégressifs. Les paramètres de l'approche proposée sont estimés en utilisant un critère généralisé d'information. Les résultats expérimentaux montrent que l'approche proposée permet de segmenter des images Térahertz avec plus de précision que d'autres approches de l'état de l'art. L'approche proposée est utilisée aussi pour détecter la nature de la surface d'un robot mobile et discriminer des événements transitoires pour assurer un fonctionnement sûr et économique du processus de surveillance.

# Chapitre 1

# État de l'art

## 1 Introduction

L'interaction du rayonnement Térahertz avec l'objet à analyser peut être définie en traitant l'ensemble de matériaux qu'ils constituent. Ces matériaux doivent avoir des réponses dans le domaine THz pour dire que les différentes structures de l'objet sont plus ou moins transmissibles ou réfléchissantes, afin de pouvoir les discriminer. L'eau et les objets humides absorbent fortement les radiations THz ; toutefois, les objets secs (tels que le papier, le tissu, le plastique et le bois) sont transparents aux radiations THz et ne fournissent pas de radiations réfléchies significatives. Les métaux sont opaques aux radiations Térahertz et reflètent la plupart des radiations entrantes. D'autres matériaux intéressants, qui offrent des radiations THz spécifiques, sont détaillés dans [54, 21].

Grace aux propriétés intéressantes des rayonnements Térahertz, plusieurs travaux d'imagerie Térahertz ont été proposés dans la littérature. Dans les travaux de Kamba et al. [69], l'imagerie Térahertz a été utilisée pour inspecter la structure des couches internes de la monture en bois sur un tableau de peinture Japonais avant sa restauration. Bowen et al. [26] ont proposé un certain nombre de techniques qui ont été utilisées pour faciliter la récupération d'images Térahertz fiables à partir d'objets complexes appartenant au domaine du patrimoine culturel. Ces techniques tentent de surmonter les problèmes posés par les surfaces inégales, en améliorant la résolution en profondeur et le contraste de l'image. Dans le domaine de la sécurité, Kowalski et al. [71] ont proposé une application qui consiste à détecter et à visualiser des objets cachés. Les propriétés des rayons THz et visibles sont exploitées et la combinaison des images fournies par les deux types de caméras permet de découvrir des objets dangereux cachés à l'intérieur des vêtements. Un certain nombre de traitements d'images Térahertz existent en littérature comprenant le débruitage d'impulsions THz, l'extraction de bandes pertinentes et la segmentation par classification de pixels THz, etc. En fait, il est connu que les systèmes d'imagerie Térahertz produisent des impulsions bruitées à cause des erreurs à la fois systématiques et aléatoires. Handley et al. [61] proposent une première méthode permettant de réduire les erreurs aléatoires. Ce travail modélise et extrait le bruit inclus dans les impulsions des images Térahertz. Ferguson [51] a proposé deux techniques principales de prétraitement : le débruitage par ondelettes et la déconvolution de Wiener. Ces méthodes ont été étudiées expérimentalement et ses performances ont été quantifiées avant de segmenter l'image Térahertz.

## 2 Travaux connexes sur la segmentation d'images Térahertz

L'image Térahertz est décrit par un nombre énorme de caractéristiques. La haute dimensionnalité des images Térahertz pose de nouveaux défis pour la détection de caractéristiques pertinentes. Le tableau 1.1 présente un résumé de plusieurs méthodes de segmentation d'images Térahertz. Certains travaux sont résumés dans cette section en termes d'espace de caractéristiques utilisées et de techniques de classification supervisées ou non supervisées. L'espace de base est constitué par les vecteurs complets dans le domaine temporel représentant les pixels de l'image Térahertz [21]. Les autres espaces de caractéristiques sont obtenus en utilisant des transformations de Fourier ou d'ondelettes [101, 110, 109]. Ces espaces peuvent être utilisés avec une seule bande ou avec plusieurs bandes. Dans le premier cas, le choix de la bande peut être fixé à priori de l'espace temporel ou spectral de l'image THz [50]. Certaines mesures issues de la forme des vecteurs représentant les pixels dans le domaine temporel ou spectral sont utilisées, telles que l'amplitude du pic maximal du vecteur [50, 22]. Dans le deuxième cas, plusieurs bandes sont utilisées, telles que le vecteur entier de l'image Térahertz, l'amplitude spectrale complète et une collection de bandes de l'image THz [50, 110, 22, 21]. Certains auteurs proposent de réduire l'espace disponible en utilisant les modèles autorégressifs, les modèles autorégressifs et moyennes mobiles, l'analyse en composantes principales et l'arbre de décision [50, 109, 113, 14, 31, 85]. La segmentation des images Térahertz est généralement réalisée en termes de classification supervisée, telles que le classificateur Mahalanobis, SVM et réseaux de neurones [50, 110, 109, 113], et de classification non supervisée, telles que $K$-means, ISODATA, hiérarchique et KHM [101, 14, 15, 22, 21, 31, 85]. Dans la section suivante, nous présentons nos contributions relatives à l'analyse d'images Térahertz.

Parmi les récents travaux, Holzinger et al. [72] ont proposé une approche de classification $k$ plus proches voisins pour segmenter les mesures Térahertz de la structure interne des dents contenant des caries. Les résultats de segmentation montrent les régions qui representent les structures internes en couches des dents. Siuly et al. [100] ont proposé une méthode d'apprentissage automatique pour la classification des images Térahertz dans le domaine biomédical. Des fonctions de corrélation croisée 2D, des méthodes d'extraction de caractéristiques statistiques et de classification standards sont utilisées ensemble pour analyser les images THz. Une étude des récents travaux de l'analyse d'images biomédicales Térahertz est détaillée dans [108].

tableau 1.1 – Résumé de quelques travaux sur la segmentation d'images Térahertz

| Methods | Features | Classification & Clustering | Application domains |
|---|---|---|---|
| Berry et al. [22] | TD & MxA & FWHM | *k*-means & ISODATA | Histopathology (basal cell carcinoma and melanoma diagnosis) |
| Berry et al. [21] | Time series & Short time Fourier transform & DWT | *k*-means | Medical (dental) & histopathology (carcinoma and melanoma diagnosis) |
| Yin et al. [109] | DWT & AR/ARMA | Mahalanobis distance classifier | Biomedical (osteosarcoma cells diagnosis) & security (mail/packaging inspection) |
| Zhong et al. [113] | PCA of the interval 0.4-1.6 THz | Minimum distance classifier & NN | Chemistry (material identification) |
| Yin et al. [110] | Spectral magnitude & spectral phase | SVM | Biomedical (ribonucleic acid recognition) & chemistry (powder identification) |
| Nakajima et al. [85] | PCA | *k*-means & AH | Histopathology |
| Brun et al. [31] | PCA | fuzzy *k*-means | Histopathology (cancer inspection from lung and pancreas) |
| Stephani [101] | DWT | Hierarchical chameleon | Security (mockup mail bomb detection) |
| Ayech et al. [14] | PCA & AR | KHM | Quality control (damage tasks detection) & agricultural (crop yield estimation) |
| Eadie et al. [50] | MxA, MnA, -MnA/MxA, MxA-MnA, FWHM, T(t), F(f), T(t)/MnA, & Decision tree | NN & SVM | Medical (colon cancer diagnosis) |

# Chapitre 2

# Segmentation d'images Térahertz utilisant $K$-means basée sur l'échantillonnage ordonné

Dans le premier chapitre de la thèse, nous proposons une nouvelle approche de segmentation d'images Térahertz basée sur la classification floue non supervisée. L'approche proposée est constituée de deux étapes. La première étape consiste à estimer les centres optimaux en utilisant une nouvelle fonction objectif basée sur l'échantillonnage d'ensembles ordonnés, alors que la deuxième étape consiste à regrouper l'ensemble de pixels observés en fonction des centres estimés. Cette approche à deux étapes est essentiellement moins sensible à l'initialisation des centres.

Dans ce chapitre, nous présentons un article intitulé **Segmentation of Terahertz imaging using $K$-means clustering based on ranked set sampling** publié dans le journal international de Elsevier **Expert Systems with Applications**, 2015 [15]. Le problème a été posé par le professeur Djemel Ziou. J'ai réalisé, validé et rédigé ce travail sous sa supervision. Une version compacte de ce travail a été publiée dans la conférence internationale **IEEE International Conference on Image Processing (ICIP2015)**, Québec, Canada, 2015, intitulée **Ranked $K$-means clustering for Terahertz image segmentation** [17].

# Segmentation of Terahertz imaging using $K$-means clustering based on ranked set sampling

Mohamed Walid Ayech
Département d'informatique, Université de Sherbrooke,
Sherbrooke, Québec, Canada J1K 2R1
walid.ayech@usherbrooke.ca

Djemel Ziou
Département d'informatique, Université de Sherbrooke,
Sherbrooke, Québec, Canada J1K 2R1
djemel.ziou@usherbrooke.ca

**Keywords**: Segmentation, Terahertz imaging, $k$-means, ranked set sampling, simple random sampling.

## Abstract

Terahertz imaging is a novel imaging modality that has been used with great potential in many applications. Due to its specific properties, the segmentation of this type of images makes possible the discrimination of diverse regions within a sample. Among many segmentation methods, $k$-means clustering is considered as one of the most popular techniques. However, it is known that $k$-means is especially sensitive to initial starting centers. In this paper, we propose an original version of $k$-means for the segmentation of Terahertz images, called ranked-$k$-means, which is essentially less sensitive to the initialization of the centers. We present the ranked set sampling design and explain how to reformulate the $k$-means technique under the ranked sample to estimate the expected centers as well as the clustering of the observed data. Our clustering approach is tested on various real Terahertz images. Experimental results show that $k$-means clustering based on ranked set sampling is more efficient than other clustering techniques such as the $k$-means based on the fundamental sampling design simple random sampling technique, the standard $k$-means and the $k$-means based on the Bradley refinement of initial centers.

# 1 Introduction

In recent years, many research groups around the world are increasing their interest on the Terahertz (THz) portion of the electromagnetic radiation [63, 112]. Terahertz radiations (T-rays) have been used in many applications, due to their interesting properties, such as noninvasive property, penetration through dry and non-metallic objects (plastic, paper, cloth, etc), and specific material characterization. The use of T-rays for imaging has opened new possibilities for research and commercial applications [54, 82, 83, 53, 69, 70, 26, 71, 65].

Terahertz pulsed imaging (TPI) system consists in collecting information from the scene, as a sequence of two-dimensional images. Each image is constituted by a set of grey level pixels acquired from a single spectral band. The combination of these images constitutes a three-dimensional Terahertz data cube. Compared to the color imaging, each pixel in the Terahertz imaging acquires many bands (e.g. 1024 bands) from the electromagnetic spectrum, instead of the only three bands of the RGB color representation. TPI system can provide specific temporal and spectral information unavailable through other sensors characterizing each pixel of the THz image. The segmentation of THz imaging supplies a wealth of information about test samples and makes possible the discrimination of heterogeneous regions within an object. Among many segmentation methods, $k$-means clustering [14, 24, 68, 77] is considered as one of the most popular techniques developed in the last few decades, due to its simplicity of implementation, fast execution and good computational performance. However, it is well known that $k$-means might converge to one of numerous local minima, and its result depends on initial starting conditions, which randomly generates the initial clustering [68]. In other words, different clustering results can be produced after different runs of $k$-means on the same input data. Given an association rule between the data points and the centers, the clustering accuracy depends on the location of the centers. The structure of the data and the sampling procedure has an effective impact on the estimation of the centers. In machine learning, the impact of sampling is often unmentioned. We show in this paper the effect of the sampling procedures in the clustering process. Simple random sampling (SRS) is the mostly used procedure in which the data points are assumed to be iid [38, 104] and there are only a few results available when the sampling design is different [20, 28, 34]. However, in some applications, such as the one explained in [77, 88, 89], using ranked set sampling (RSS), may be cheaper and result in better and more informative samples from the underlying population. In this paper, we study the problem of initial center sensitivity of $k$-means technique; explain how to reformulate the $k$-means under the RSS design to overcome the initialization problem and classify the observed data. The obtained results are compared with the corresponding ones of simple random sample data. We show that, using RSS, our approach leads to a better inference about the precision of centers and therefore the precision of the obtained clusters. Experimental tests of our approach are done to segment Terahertz images. The obtained results show the interest of ranking the pixels and explain how the extra information via the rank of each pixel in RSS will lead to a more efficient classification of pixels compared with SRS and other techniques.

The rest of the paper is organized as follows: in section 2, we give an insight about related works of various imaging applications in the Terahertz domain. Section 3 introduces the $k$-means clustering based on the simplest sampling design SRS technique that we call SRS-$k$-means. Section 4 presents the RSS technique and explains its efficiency compared to the SRS. The formulation of the general $k$-means in the case of RSS sample and the different steps of the resulting algorithm, ranked-$k$-means, are also described. Our clustering approach based on RSS sample is compared with the clustering approach in the case of SRS, the standard approach of $k$-means and the $k$-means using the Bradley refinement of initial centers on the real Terahertz images of a carbon fiber sample, a flexure spring and a fruit grape. The results are illustrated and discussed in section 5.

## 2 Related works on Terahertz image segmentation

The Terahertz image is formed by capturing THz radiations reflected from or transmitted through objects. Water and moisture objects highly absorb THz radiations, however, dry objects (such as paper, cloth, plastic and wood) are transparent to THz radiations and provide no significant reflected radiations. Metals are opaque to THz radiations and reflect most incoming radiations. Other interesting materials, which offer specific THz radiations, are detailed in [54, 21]. The THz image is formed by several bands (e.g. 1024 bands). The high dimensionality of THz images leads to some new challenges for relevant feature detection. Indeed, the relevant features can be embedded only on few bands [54, 21]. For this raison, in several related works, the band having the maximal pick amplitude is used. It has pointed out that other bands may contains relevant features and these bands are not known in advance [14, 109, 85]. The features are used for the segmentation of THz images. In the most related works, classification of features is used for the segmentation of Terahertz images.

Table 2.1 presents a summary of several segmentation methods, regrouped in terms of feature space used, classification or clustering techniques and application domains. From this table, we deduce three important remarks. The first one concerns the various application domains using the Terahertz imaging which explains the interest of analysing this now technology of imaging. The second remark concerns the feature spaces used in the state of art. The basic feature space is the raw time series of THz images [21]. Other feature spaces are obtained by using Fourier or Wavelet transforms [109, 110, 101]. The feature space can be used with only one band or with several bands. In the first case, the choice of the band can be priori fixed either from the time series space (T(time=constant)) or from the spectral space (F(frequency=constant)) of the THz image [50]. Some measures from the shape of the entire time series or other spectral transform are used, such as the maximal pick amplitude of the time series (MxA), the minimal pick amplitude of the time series (MnA), the time delay (TD) of the maximal pick of the time series, the full width at half maximum pick (FWHM) [22, 50]. In the second case, several bands are used, such as the full time series of the THz image, the full spectral amplitude, the full spectral phase, and a collection of some bands such as MxA, MnX and FWHM [21, 110, 50, 22]. To reduce the feature space, autoregressive model (AR), autoregressive moving average model (ARMA), principal component analysis (PCA) and decision tree are used as extraction or selection methods [109, 113, 85, 31, 14, 50]. The third remark concerns methods of segmentation allowing identification of different regions in Terahertz images. These methods are supervised (classification), such as Mahalanobis distance classifier, minimum distance classifier, support vector machine (SVM) and neural networks (NN) [109, 113, 110, 50], and unsupervised classification, such as hard $k$-means,

Table 2.1 – Summary of some related works on Terahertz image segmentation

| Methods | Berry et al. [22] | Berry et al. [21] | Yin et al. [109] | Zhong et al. [113] | Yin et al. [110] | Nakajima et al. [85] | Brun et al. [31] | Stephani [101] | Ayech et al. [14] | Eadie et al. [50] |
|---|---|---|---|---|---|---|---|---|---|---|
| Features | TD & MxA & FWHM | Time series & Short time Fourier transform & DWT | DWT & AR/ARMA | PCA of the interval 0.4-1.6 THz | Spectral magnitude & spectral phase | PCA | PCA | DWT | PCA & AR | MxA, MnA, -MnA/MxA, MxA-MnA, FWHM, T(t), F(f), T(t)/MnA, & Decision tree |
| Classification & Clustering | *k*-means & ISO-DATA | *k*-means | Mahalanobis distance classifier | Minimum distance classifier & NN | SVM | *k*-means & AH | fuzzy *k*-means | Hierarchical chameleon | KHM | NN & SVM |
| Application domains | Histopathology (basal cell carcinoma and melanoma diagnosis) | Medical (dental) & histopathology(carcinoma and melanoma diagnosis) | Biomedical (osteosarcoma cells diagnosis) & security (mail/packaging inspection) | Chemistry (material identification) | Biomedical (ribonucleic acid recognition) & chemistry (powder identification) | Histopathology | Histopathology (cancer inspection from lung and pancreas) | Security (mockup mail bomb detection) | Quality control (damage tasks detection) & agricultural (crop yield estimation) | Medical (colon cancer diagnosis) |

fuzzy $k$-means, ISODATA, hierarchical chameleon, agglomerative hierarchical (AH) and k-harmonic-means (KHM) [22, 85, 21, 31, 101, 14]. In our previous work [14], the combined AR/PCA model are used to extract relevant features from the high dimensional THz images of a carbon fiber sample, a flexure spring and a fruit grape. These features are used in this paper for the validation purpose.

The $k$-means clustering has been shown efficient for the segmentation of Terahertz images [22, 21, 31, 85]. However, $k$-means techniques are especially sensitive to initial starting conditions and different runs of $k$-means segmentation on the same input Terahertz image can produce different results. In the literature of $k$-means algorithm, the sampling procedures is not exploited since the whole observed dataset, i.e. all the pixels of the image, is used in the clustering process. We show in this paper the effect of the sampling procedures in the clustering accuracy of the $k$-means technique. Ranked set sampling procedure is proposed to extract a representative sample from the observed population and provide therefore conclusions about the centers. Our approach which is called ranked-$k$-means consists to reformulate the standard $k$-means under the ranked sample to overcome the initialization problem and classify the observed Terahertz data. Ranked-$k$-means is compared with the corresponding ones using the simple random sampling that is called SRS-$k$-means and presented in section 3.

# 3 SRS-$k$-means clustering

In data clustering, it is typically assumed that the data point observations, such as the pixels of the THz image, are drawn from the continuous populations which correspond to natural phenomena such as the real scene before image acquisition. Obviously, the observed populations which constitute the accessible part from the continuous population are only studied in order to attempt to learn something about the inaccessible population. In this paper, the term "observed population" is simplified to "population". It is constituted by a set of $N$ data point realizations, denoted $X$ and represented as follows $\{x_1, ..., x_N\}$. We are interested in this section to classify the observed dataset into homogeneous clusters. One of the well known clustering techniques, $k$-means [77, 24], is used to regroup the dataset into a pre-defined number

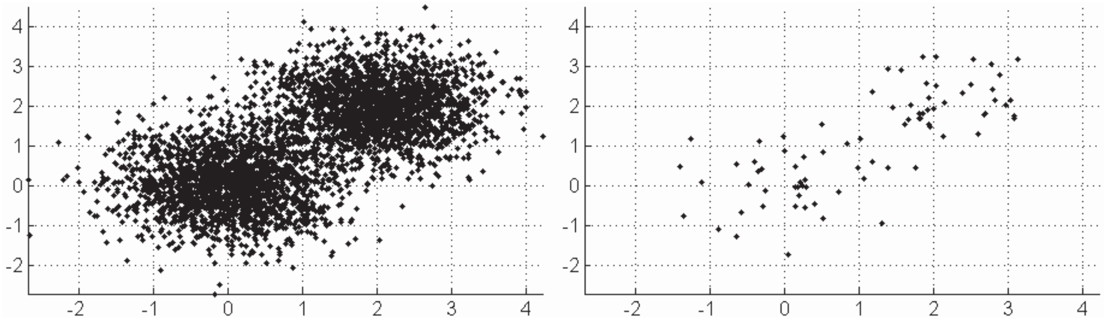


Figure 2.1 – Example of data point observations distributed in bi-dimensional subspace: the full dataset ($N$=4000) in the left versus a small simple random sample points ($n$=40) in the right.

$L$ of clusters. $K$-means begins by randomly choosing $L$ centers from the studied dataset, one center for each cluster. $K$-means is a two steps iterative algorithm. The first step consists to assign each data point to the cluster having the closest center using either hard or soft decision rule [77, 24]. The second step consists to update the centers values by computing the weighted average of data points belonging to each cluster. The centers adjust their locations behind every iteration using the whole observed data points until convergence, i.e. when the centers do not change. Making use of the whole observed data points into the clustering process, the behavior of the standard $k$-means can be considered as an exhaustive analysis, in other words, the clustering of the whole unit measurements may occupy high time and memory consuming, while a small number of sampling units can be practically fast, easily and accurately representative of the whole observed population. In the state of art, few authors used random sampling in order to avoid the use of the whole set of available data [77]. Among them, Bejarano et al. [20] have proposed a sampling method that consists to randomly permute the data point observations; follow by only extracting the first $n$ points from the permuted dataset as input of the $k$-means technique to decrease runtime. Bradley and Fayyad [28] have proposed other approach to estimate initial conditions for the $k$-means clustering. This approach is based on multiple small random sub-sample solutions ($T$ sub-samples). These sub-samples are classified using $k$-means randomly initialized producing $T$ sets of intermediate centers, each one with

$L$ points. These center sets are fused into a superset containing $T \times L$ points. The superset is then classified by $k$-means $T$ times, each time initialized with a different center set. Centers having the minimal objective function value are selected and then considered as refined initial points for the clustering process. The sampling methods used by Bejarano and Bradley are proprietary in the sense that they seems not be known in statistics. Therefore, even if they perform well in the experimentations carried by authors, their properties are unknown.

In this section, we propose an issue which seems not has been tackled concerning the combination of the widely studied probabilistic sampling design: simple random sampling (SRS) [38, 104] and the $k$-means clustering technique. A representative sample from the observed population is then randomly selected and regrouped into homogeneous clusters in order to get conclusions about the centers. In SRS, the studied population, which represents the observed dataset $X$, must be firstly defined and a good choice of the sample size $n$ requires to be fixed. The size of the sample must be chosen so as to achieve the best tradeoff between the estimation accuracy and the low computational cost. In the SRS process the data are assumed to be i.i.d. and each sample may be drawn either with or without replacement. In our work, the sample is drawn with replacement, so each element from the population list is selected and therefore returned to the list to be able selected another time. SRS with replacement are used to fortify data points having the maximum chance of being selected and offer therefore more reliability to the drawn sample. The main steps necessary to selecting a simple random sample data are summarized as follows:

```
SRS(X,n) algorithm
1. Develop a population list of all the elements of the studied population and
assign each element a number to be able to access to the population.
2. Generate a list of $n$ random numbers.
3. Select the elements $\{x_1, ..., x_n\}$ that have numbers corresponding to the
generated random number list and save them in a dataset denoted $X_{SRS}$.
4. Return $X_{SRS}$.
```

Using the procedure SRS($X$,$n$), a random sample of $n$ points is collected from the observed population $X$ and stored as a new small dataset $X_{SRS} = \{x_1, ..., x_n\}$. The obtained sample represents the data under study and allows providing inferential statistics for the whole observed data clustering. Figure 2.1 shows an example of a set of data point observations $X$ distributed in two clusters. A small random sample $X_{SRS}$ is drawn from the observed population. Both $X$ and $X_{SRS}$ are shown respectively on the left and on the right. Each point on the right may be considered as a possible point from the observed population having a random location. The datasets $X$ and $X_{SRS}$ have sizes respectively equal to 4000 and 40. The SRS sample represents only 1% of the observed population. The possible $X_{SRS}$ samples are comparatively varied; however they can produce an expected behavior, only needed in our work, to give an estimate about the centers. Then, a data sample $X_{SRS}$ is used as input for the $k$-means algorithm. The novel version of $k$-means based on SRS is called SRS-$k$-means algorithm. When the algorithm converges, we have a small random sample from the population that has been regrouped into $L$ clusters. On the basis of the obtained centers, all the $N$ points in the observed dataset are now classified by affecting each observed data point to the nearest cluster represented by its center. The $L$ obtained clusters are the final output of our clustering based random sampling algorithm. As discussed above, SRS-$k$-means is a two-step algorithm. The first step is called E-step (estimation step) and consists to classify the $X_{SRS}$ data into $L$ clusters; each one is represented by its own estimated center $m_l$. The E-step of SRS-$k$-means consists to minimize the objective function which is defined as follows:

$$JE_{SRS} = \sum_{l=1}^{L} \sum_{j=1}^{n} u_{jl}^{a} d(x_j, m_l), \tag{2.1}$$

where $u_{jl}$ represents the membership degree of the $j^{th}$ object in the $l^{th}$ cluster; $d$ represents a distance metric (generally the Euclidian distance) that measures the similarity between an object and a cluster center, and $a > 1$ represents the degree of fuzzification. The values of the membership function $u_{jl}$ must verify the constraints $\{u_{jl} \mid u_{jl} \in [0, 1] \, \text{and} \sum_{l=1}^{L} u_{jl} = 1\}$. Then, the first order condition allows writing the membership degrees and the centers as follows:

$$u_{jl} = \left( \sum_{h=1}^{L} \left( \frac{d(x_j, m_l)}{d(x_j, m_h)} \right)^{\frac{1}{a-1}} \right)^{-1} \tag{2.2}$$

and

$$m_l = \frac{\sum_{j=1}^{n} u_{jl}^{a} x_j}{\sum_{j=1}^{n} u_{jl}^{a}}. \tag{2.3}$$

Clustering step (C-step) is the second step of SRS-$k$-means and consists to affect each point from the observed data $X$ to the nearest cluster represented by its estimated center. For that reason, we propose to estimate the membership degree of each data point by minimizing the objective function given by the equation (2.1) using the size of the whole dataset $N$ instead of the sample size $n$. We note that this objective function can be also used in the case where only the $n$ data sample points are bringing into the C-step. The first order condition allows computing the membership degrees of data points to clusters as given by the equation (2.2).

The SRS-$k$-means technique appears faster than the traditional $k$-means; however, the sensitivity problem to initial centers remains not yet resolved. To overcome this drawback, we propose an extended approach of the SRS-$k$-means called ranked-$k$-means which is implemented based on a more sophisticated sampling design ranked set sampling.

# 4 Ranked-$k$-means clustering

The sampling methods can have a great influence on the performance of the $k$-means clustering since classifying the full measurement of the variable of interest is assumed costly. Considering that the cost of drawing a sample and ranking is negligible, ranked set sampling (RSS), which is used in some applications, such as the one explained in [80, 88, 89, 107], may provide cheaper, better and more informative samples from the underlying population. Compared to SRS, RSS has been proven theoretically [103] and shown empirically [37] to yield more precise estimator of the population mean. Moreover, it has been shown that RSS provides an extra special design structure that can be used to improve many data mining applications [88]. In this paper, we propose a novel approach of clustering called ranked-$k$-means that consists to reformulate the $k$-means algorithm by using the RSS design. In RSS, a set of data point realizations of some variables of interest is drawn by a SRS procedure and then the data points in the set are ranked according to some pre-specified criterion.

In one dimensional case, it is obvious that data points can be ranked by using the values of the variable of interest. Nevertheless, ranking multidimensional data sample remains more complex. In the original version of the RSS [80], the ranking is done by human judgment. However, for large datasets, only few applications exist in the literature of RSS where ranking can be carried out by judgment with respect to the variable of interest. In other versions of the RSS [88, 89], ranking is performed by using some covariate information which can be represented as another available variable, called concomitant variable. The concomitant variable, denoted $Y$, must be highly correlated with the main variable of interest, but requires negligible cost. We adapt this ranking criterion in our work to sort the different units in the RSS process. Since then, different ranking procedures have been devised for different purposes such as in [88, 89]. All the RSS variants share the same basic features and properties. Samples ascertained through the RSS procedure contain more information (according to Fisher information) than SRS of the same size, which explain why RSS is more efficient than SRS as has demonstrated by many previous works [80, 89]. For our purpose, let us consider a set $X$ of $N$ observed data points and a set $X_{RSS}$ of $n = R \times K$ data sample points. The data sample points of a cycle $r$ are generated as follows. A $K \times K$ samples are randomly drawn from the set $X$; i.e. SRS sampling. The sample points $\{x_{i1}, \ldots, x_{iK}\}$ of the $i^{th}$ row are ranked; that is $x_{i(1)} \leq \ldots \leq x_{i(K)}$. The $i^{th}$ minimum is saved. All the kept minima form the $r^{th}$ cycle $\{x_{r1}, \ldots, x_{rK}\}$. The procedure is repeated $R$ times (i.e., $R$ cycles) and constitutes the ranked sample $X_{RSS}$ of dimension $R \times K$. The RSS algorithm can be therefore summarized as follows:

```
*RSS*($X$,$K$,$R$) *algorithm*
**1. for** $r \leftarrow 1$ to $R$ **do**
        $X_{SRS} \leftarrow$ SRS($X$,$K^2$)
        Reformulate the $X_{SRS}$ in the $K \times K$ matrix form $A$.
        Sort each $k^{th}$ row of $A$ according to a concomitant variable $Y$ and access
          the $k^{th}$ minimum.
        Store the resulting minima $(x_{r1}, \ldots, x_{rK})^t$ in the $r^{th}$ row of the matrix
          $X_{RSS}$ of dimension $R \times K$.
    **end**
**2.** Return $X_{RSS}$
```

Regrouping both samples $X_{RSS}$ and $X_{SRS}$ into $L$ clusters is expected to give more precise centers $m_{RSS}$ than $m_{SRS}$ ($l = 1 \dots L$) when both samples are based on the same sample size. In the case where $L$ equal to 1, only one cluster is used, and the centers $m_{1,RSS}$ and the centers $m_{1,SRS}$ represent also the means respectively of $X_{RSS}$ and $X_{SRS}$ samples. In this case, McIntyre [80] has stated a relationship between SRS and RSS which is defined as follows:

$$var(\hat{m}_{1,RSS}) = var(\hat{m}_{1,SRS}) - \frac{1}{RK^2}\sum_{k=1}^{K}(w_k - w)^2 \tag{2.4}$$

where $w_k = \sum_{j=1}^{R} x_{jk}$ and $\sum_{k=1}^{K} w_k = Kw$. The inequality (2.5) can be therefore easily deduced as follows:

$$var(\hat{m}_{1,RSS}) \leq var(\hat{m}_{1,SRS}) \tag{2.5}$$

The variance of $m_{1,RSS}$ is always less than or equal to the variance of $m_{1,SRS}$ regardless of ranking errors which confirm the precision of the RSS regarding to the SRS in the case of $L = 1$. When $K = 1$, the sample size $n$ becomes equal to $R$, and SRS and RSS samples give the same variance.

The inequality of equation (2.5) can be generalized in the case where $L \geq 1$. The new relationship is then deduced as follows:

$$\sum_{l=1}^{L} var(\hat{m}_{l,RSS}) \leq \sum_{l=1}^{L} var(\hat{m}_{l,SRS}) \tag{2.6}$$

For $l = 1 \dots L$, the sum of variances of $\hat{m}_{l,RSS}$ is always less than or equal to the correspondent of $\hat{m}_{l,SRS}$. In fact, the precision of the $\hat{m}_{l,RSS}$ centers return essentially to the reformulation of the population into spaced and compact small subpopulations. Figure 2.2 shows the same example of the data point realizations in Figure 2.1. The full dataset $X$ and a small RSS sample $X_{RSS}$ are shown respectively on the left and on the right. In this example, the $X_{RSS}$ has the same size of the $X_{SRS}$ in Figure 2.1. The RSS population is then formulated into $K$ ranked subpopulations ($K$=2 in the example of Figure 2.2) where each one has its own distribution (first and second subpopulations are symbolized respectively by "o" and "+"). The structure of the ranked sample reflects the decrease of the ambiguity between subpopulations. In the general case where the clusters number $L \geq 2$, the obtained RSS measurements are

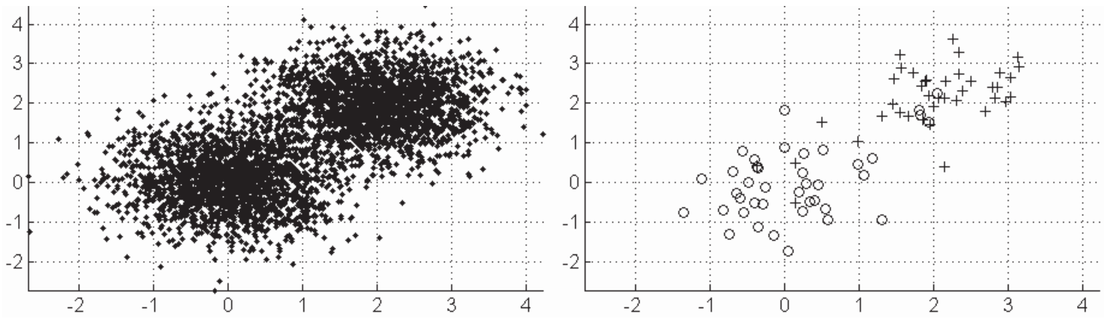


Figure 2.2 – Example of data point observations distributed in bi-dimensional subspace: the full dataset ($N$=4000) in the left versus a small ranked set sample points ($n$=40 and $K$=2) in the right.

expected to be more regularly spaced than those obtained through SRS and more representative of the underlying population.

Let us consider the parameter $w_k$ which represents the mean of the $k^{th}$ subpopulation of $X_{RSS}$, given by the following equation:

$$w_k = \sum_{r=1}^{R} x_{rk} \tag{2.7}$$

In addition to the ranked sample $X_{RSS}$, the parameter $w_k$ is used in our clustering approach to incorporate the rank information into the $k$-means process. In similar way to the SRS-$k$-means, ranked-$k$-means is also a two-step algorithm. The E-step consists to classify the $X_{RSS}$ data sample into $L$ clusters, each one is represented by its own estimated center $\hat{m}_l$. In the E-step, an objective function of the ranked-$k$-means will be penalized by a regularization term, which is integrated in the distance measure. Let us consider the following parameters: $d$ represents a distance metric (generally the Euclidian distance), $m_l$ represents the center of the $l^{th}$ cluster, $u_{rkl}$ represents the membership degree of $x_{rk}$ in the $l^{th}$ cluster, and $a > 1$ represents the degree of fuzzification. The new objective function is defined as follows:

$$JE_{RSS} = \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{r=1}^{R} u_{rkl}^{a} [d(x_{rk}, m_l) + \alpha \times d(w_k, m_l)], \tag{2.8}$$

Where parameter $\alpha$, in the second term, controls the effect of the order statistic of $x_{rk}$. In essence, the addition of the second term in equation 2.8 formulates a rank constraint and aims at keeping the closeness of the prototype of the $l^{th}$ cluster around the nearest subpopulation means $w_k$ computed from the RSS sample. By an optimization way, the objective function $JE_{RSS}$ can be minimized with respect to $u_{rkl}$ and $m_l$. The values of the membership function $u_{rkl}$ must verify the constraints $U = \{u_{rkl} | u_{rkl} \in [0,1]$ and $\sum_{l=1}^{L} u_{rkl} = 1\}$. The membership degree of $x_{rk}$ to the $l^{th}$ cluster and the center $m_l$ of the $l^{th}$ cluster are given as follows:

$$u_{rkl} = \left( \sum_{h=1}^{L} \left( \frac{d(x_{rk}, m_l) + \alpha \times d(w_k, m_l)}{d(x_{rk}, m_h) + \alpha \times d(w_k, m_h)} \right)^{\frac{1}{a-1}} \right)^{-1} \tag{2.9}$$

and

$$m_l = \frac{\sum_{k=1}^{K} \sum_{r=1}^{R} u_{rkl}^{a} (x_{rk} + \alpha \times w_k)}{(\alpha + 1) \sum_{k=1}^{K} \sum_{r=1}^{R} u_{rkl}^{a}}. \tag{2.10}$$

When $\alpha$ is set to zero, the algorithm is equivalent to the original $k$-means applied on the $X_{RSS}$ sample, while increasing the value of $\alpha$, the algorithm promotes the rank effect for each $x_{rk}$ to its own subpopulation in RSS sample.

The ranked-$k$-means requires therefore reorganizing the observed data $X$ in subpopulation form ($X_{RSSF}$) to be used in the C-step process. The order statistic $k(x_j)$ of each data point $x_j$ is then estimated by minimizing the equation $\sum_{k=1}^{K} \sum_{r=1}^{R} (d(x_j, x_{rk}) + \alpha d(x_j, w_k))$ with respect to the index $k$. The obtained dataset is then rearranged in a $R_k \times K$ matrix, where $R_k$ represents the entire cycle number of each order statistic in $X_{RSSF}$. However, the minimized equation may occupy high time consuming, that is why we propose to reduce the equation as follows $\sum_{k=1}^{K} d(x_j, w_k)$ which gives approximately the same results. The C-step consists therefore to regroup each element from the $X_{RSSF}$ to the nearest cluster represented by its own estimated center. For that reason, we propose to estimate the membership degree of each observed point $x_{rk}$ to the $l^{th}$ cluster using the following objective function:

$$JC_{RSS} = \sum_{k=1}^{K} JC_{RSS,k}, \tag{2.11}$$

where

$$JC_{RSS,k} = \sum_{r=1}^{R_k} \sum_{l=1}^{L} u_{rkl}^{a} (d(x_{rk}, m_l) + \alpha \times d(w_k, m_l)). \tag{2.12}$$

By an optimization way, the objective function $JC_{RSS}$ can be minimized with respect to the $u_{rkl}$. The membership degrees of data points to clusters are then given by the equation (2.9). The order statistic of each data point can be defined as follows:

$$k = \arg\min_{k}(JC_{RSS,k}) \tag{2.13}$$

The two steps which constitute the ranked-$k$-means algorithm can be summarized as follows:

```
E-step algorithm
1. Data: $X_{RSS}$ = RSS($X$,$K$,$R$) algorithm
   Compute $w_k$ as the mean of the $k^{th}$ subpopulation
   Initialize the centers $m_l$ by random points from $X_{RSS}$
2. Do
      Update centers $m_l$ using Eq. (2.10)
      Update membership degrees $U$ using Eq. (2.9)
   Until $|m^t - m^{t-1}| <$ threshold1
3. Return the obtained centers $m_l$
```

```
C-step algorithm
1. Data: $X_{RSSF}$ (rearranged data in $R_k \times K$ matrix)
   Centers: $m$ (centers obtained from the E-step)
2. Do
      Update the membership degree $U$ using Eq. (2.9)
      Update the order statistic using Eq. (2.13)
   Until $|U^t - U^{t-1}| <$ threshold2
3. Return $U$.
```

In the E-step, the $X_{RSS}$ sample is obtained by applying the RSS($X$,$K$,$R$) algorithm on the observed data $X$. The output of the E-step is constituted by the estimated centers. In the C-step, the membership degrees and the order statistics of data points adjust their values behind every iteration until the membership degrees do not change.

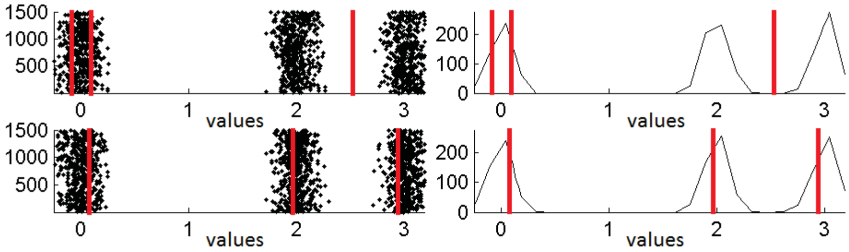


Figure 2.3 – Dataset 1 constituted by three Gaussian distributions. First and second rows represent respectively SRS and RSS samples with its histograms. Ranked-$k$-means is used with $\alpha$=0.2 and $K$=3. Red vertical lines represent the centers of each cluster obtained by the E-step of SRS and ranked $k$-means.

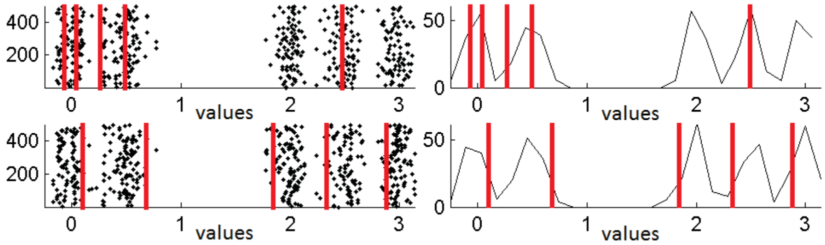


Figure 2.4 – Dataset 2 constituted by five Gaussian distributions. First and second rows represent respectively SRS and RSS samples with its histograms. Ranked-$k$-means is used with $\alpha$=0.4 and $K$=5. Red vertical lines represent the centers of each cluster obtained by the E-step of SRS and ranked $k$-means.

# 5 Experimental results

In this section, we investigate the performance of our approach on real Terahertz images. As the ground truth of the THz images is not very precise in our work, we start firstly by studying and showing the performances of the clustering techniques on several artificial and real standard datasets. Since the clustering accuracy depends on the location of the centers, we experimentally show in this case that SRS has ineffective impacts. In this section, we empirically show that by incorporating the rank information about the sample points, ranked set sample can be more representative of the true underlying population and therefore more efficient to estimate centers for the process of the clustering step. Experimental tests have been realized on artificial, standard and Terahertz datasets.

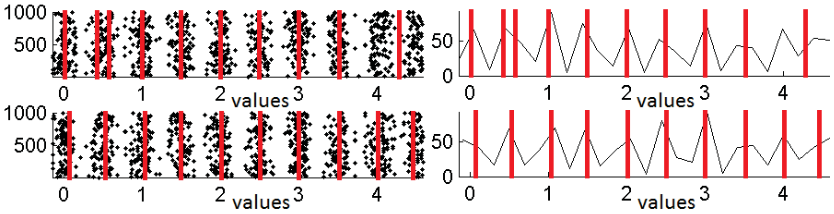


Figure 2.5 – Dataset 3 constituted by ten Gaussian distributions. First and second rows represent respectively SRS and RSS samples with its histograms. Ranked-$k$-means is used with $\alpha$=0.2 and $K$=10. Red vertical lines represent the centers of each cluster obtained by the E-step of SRS and ranked $k$-means.

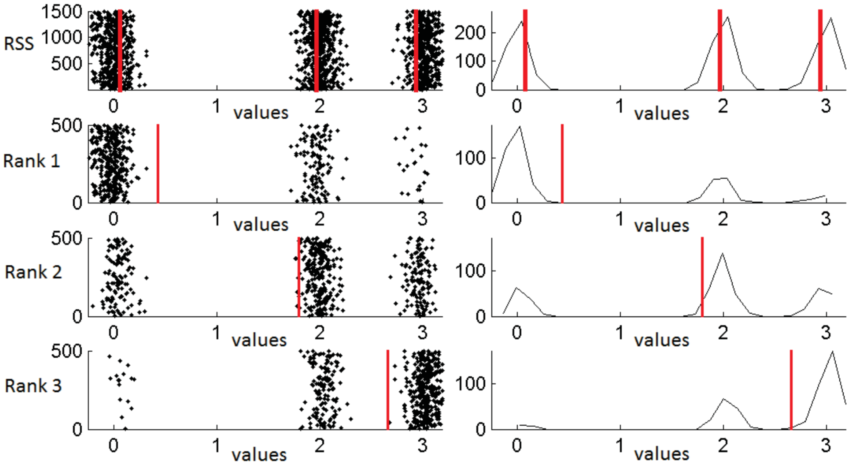


Figure 2.6 – First row represents RSS sample of dataset 1 and its histogram, and the red vertical lines represent the estimated centers $m_l$. Below the first row, each row represents one subpopulation of RSS and its histogram, and the red vertical lines represent the subpopulation means $w_k$.

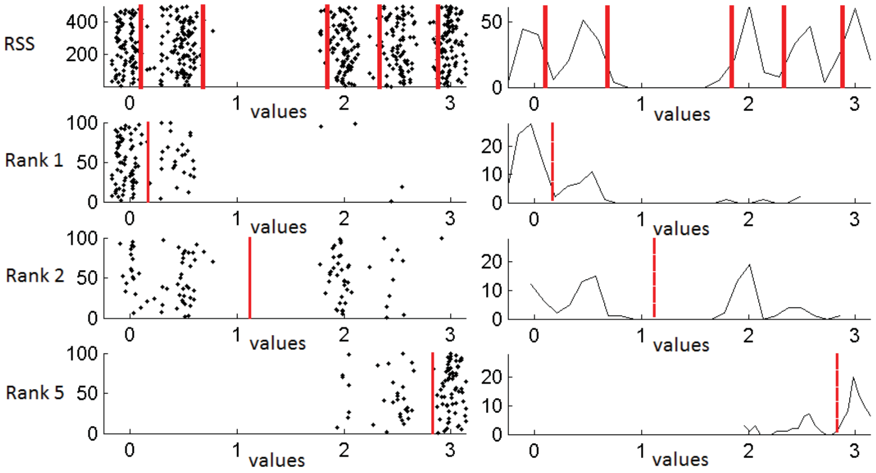


Figure 2.7 – First row represents RSS sample of dataset 2 and its histogram, and the red vertical lines represent the estimated centers $m_l$. Below the first row, each row represents one subpopulation of RSS and its histogram (for ranks 1, 2 and 5), and the red vertical lines represent the subpopulation means $w_k$.

## 5.1 Artificial and standard datasets

In order to study with more details the performance of our approach, we start our tests on mono-dimensional datasets. This choice enabled us to simply and strongly show the interests of our approach and offer better illustration of the results. Interpretations that will be acquired by this study will then be facing more complex situation of multi-dimensional Terahertz data in section 5.2. Three artificial datasets were generated respectively by three, five and ten Gaussian distributions and comport respectively 1500, 500 and 1000 points. Our approach based on RSS is compared to SRS-$k$-means. Figures from 2.3 to 2.5 show the used datasets with both samples. For each dataset in these figures, first and second rows represent respectively SRS and RSS samples in the left and its histograms in the right. Experimental tests are done in the same condition that is the same initial centers. RSS design was used with a cycle number $R$ equal to the round of $N/K$. For datasets from 1 to 3, RSS was used respectively with $K$ equal 3, 5, and 10, i.e. with $K = L$. The parameter $\alpha$ is fixed to 0.2 for dataset 1 and dataset 3, and 0.4 for dataset 2. It is shown in figures 2.3, 2.4 and 2.5 that ranked samples are more effective than random samples to estimate centers (represented by the red vertical lines) for the C-step process. Obviously, fig-

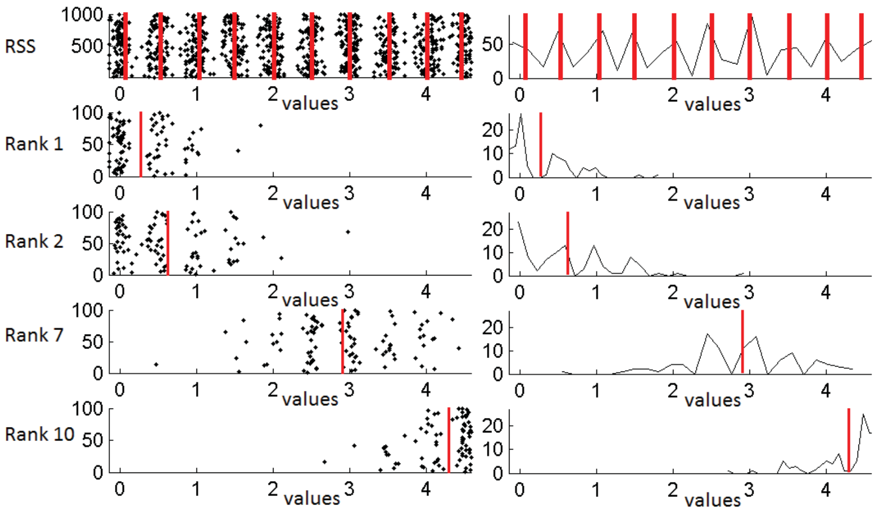


Figure 2.8 – First row represents RSS sample of dataset 3 and its histogram, and the red vertical lines represent the estimated centers $m_l$. Below the first row, each row represents one subpopulation of RSS and its histogram (for ranks 1, 2, 7 and 10), and the red vertical lines represent the subpopulation means $w_k$.

ure 2.3 shows that for SRS sample, the center locations of all clusters are skewed. On the other side, RSS gives better estimation of the three centers. The same thing in figures 2.4 and 2.5, the centers of the $2^{nd}$ and $3^{rd}$ clusters of SRS sample of dataset 2 and the centers of $2^{nd}$, $9^{th}$ and $10^{th}$ clusters of SRS sample of dataset 3 are clearly skewed. While, in RSS samples, the centers are better located and show that RSS appear more representative of the population mean of different clusters of dataset 2 and dataset 3. In fact, the RSS sampling procedure consists to reformulate the datasets into $K$ ranked sub-sets where each one has its own distribution visualized below the first row of figures 2.6, 2.7 and 2.8. This reformulation reflects the decrease of the ambiguity between the sub-sets. The obtained measurements of RSS are expected to be more regularly spaced than those obtained through SRS, and consequently ranked sample is more representative of the whole population.

In this study, our estimation approach depends on four parameters: the fuzzification degree $a$, the rank effect $\alpha$, the subpopulation number $K$ and the sample size $n$. The sample size $n$ and the cycle number $R$ are strictly dependent because $n$ is equal to $R \times K$. Our tests was applied for all datasets with $a = 2$, as that was one

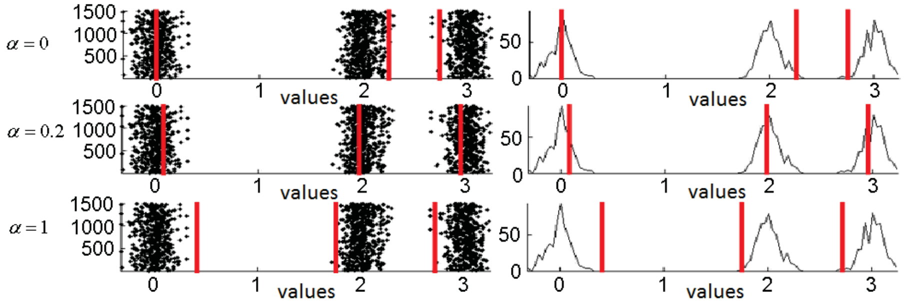
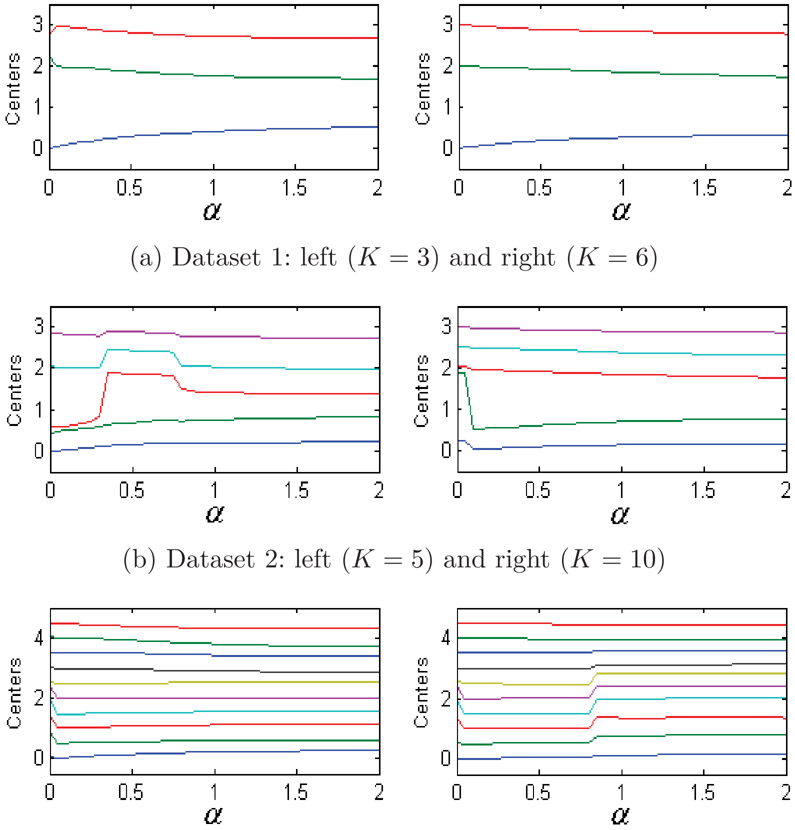


Figure 2.9 – Center positions obtained by our estimation approach in dataset 1 for $\alpha$=0, $\alpha$=0.2 and $\alpha$=1.

of the best values used in the literature of fuzzy clustering techniques [24]. In the rest of this section, we will study the effect of $\alpha$, $K$ and $n$ on the achieved tests. Figure 2.9 shows the center positions obtained by our estimation approach on dataset 1 for three different values of $\alpha$ (0, 0.2 and 1). In this figure, we observe, for $\alpha = 0$, that the centers of the $2^{nd}$ and $3^{rd}$ clusters are clearly skewed. On the other hand, for $\alpha = 0.2$, the centers are clearly better located. For $\alpha = 1$, the centers come close to the mean $w_k$ of each of the three subpopulations in figure 2.6. Figure 2.10 shows the variation of the found centers of clusters for each real value of $\alpha$ from 0 to 2. In this figure, tests are achieved for the three Gaussian datasets with two different values of $K$. We observe that for $\alpha$ equal to 0, the centers of the $2^{nd}$ and $3^{rd}$ clusters are clearly skewed in figure 2.10 (a), $3^{rd}$ center (for $K = 5$), $1^{st}$ and $2^{nd}$ centers (for $K = 10$) in figure 2.10 (b) are largely skewed. Furthermore, in figure 2.10 (c), the centers of $2^{nd}$, $3^{rd}$, $4^{th}$ and $5^{th}$ clusters (for $K = 10$) and $3^{rd}$, $4^{th}$ and $5^{th}$ centers (for $K = 20$) are skewed. Incrementing $\alpha$ between 0.2 and 1, all the centers became better located for datasets. Beyond the value 1, the centers become stationeries and close to the subpopulation mean $w_k$, represented by the red vertical lines in figures 2.6, 2.7 and 2.8 of the $K$ subpopulations of dataset 1, dataset 2 and dataset 3. Figure 2.11 shows the effect of parameter $K$ on the obtained center positions of dataset 1. In this figure, we observe, for $K = 1$, that the first center is inclined to the right, and, the $2^{nd}$ and $3^{rd}$ centers are fused in the position 0.4. On the other hand, for $K = 3$, the centers are well located with a slightly deviation of the first center. For $K = 6$, the centers are perfectly located for the three clusters. Figure 2.12 shows the variation



(a) Dataset 1: left ($K = 3$) and right ($K = 6$)

(b) Dataset 2: left ($K = 5$) and right ($K = 10$)

(c) Dataset 3: left ($K = 10$) and right ($K = 20$)

Figure 2.10 – Variation of the obtained centers for each value of $\alpha$ (from 0 to 2) for the three datasets. The parameter $K$ is equal respectively to 3, 5 and 10 in the left and 6, 10 and 20 in the right. The sample size $n$ is fixed to the whole dataset size $N$.

of the found centers of clusters for each value of $K$ from 1 to 20. In this figure, tests are achieved for the three Gaussian datasets with two different values of $\alpha$. For all datasets, we observe that for the lowest values of $K$, the centers are very poorly located, such as the centers of the $2^{nd}$ and $3^{rd}$ clusters in figure 2.12 (b), $2^{nd}$ clusters (for $K = 5$) and $2^{nd}$ and $3^{rd}$ clusters (for $K = 10$) in figure 2.12 (c), $2^{nd}$, $3^{rd}$, $4^{th}$ and $5^{th}$ clusters (for $K = 10$) and $2^{nd}$, $3^{rd}$ and $4^{th}$ clusters (for $K = 20$) in figure 2.12 (d). Increasing the value of $K$, all center locations are highly improved essentially when $K$ exceeds $L$. We observe that for a good choice of $\alpha$ and $K$, our estimation approach appears more robust and significant centers can be therefore estimated for the clus-

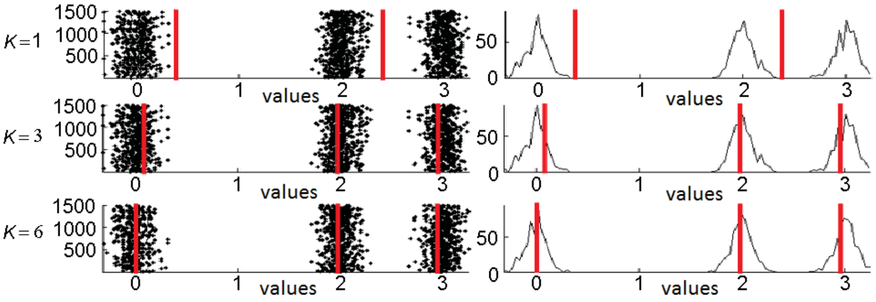
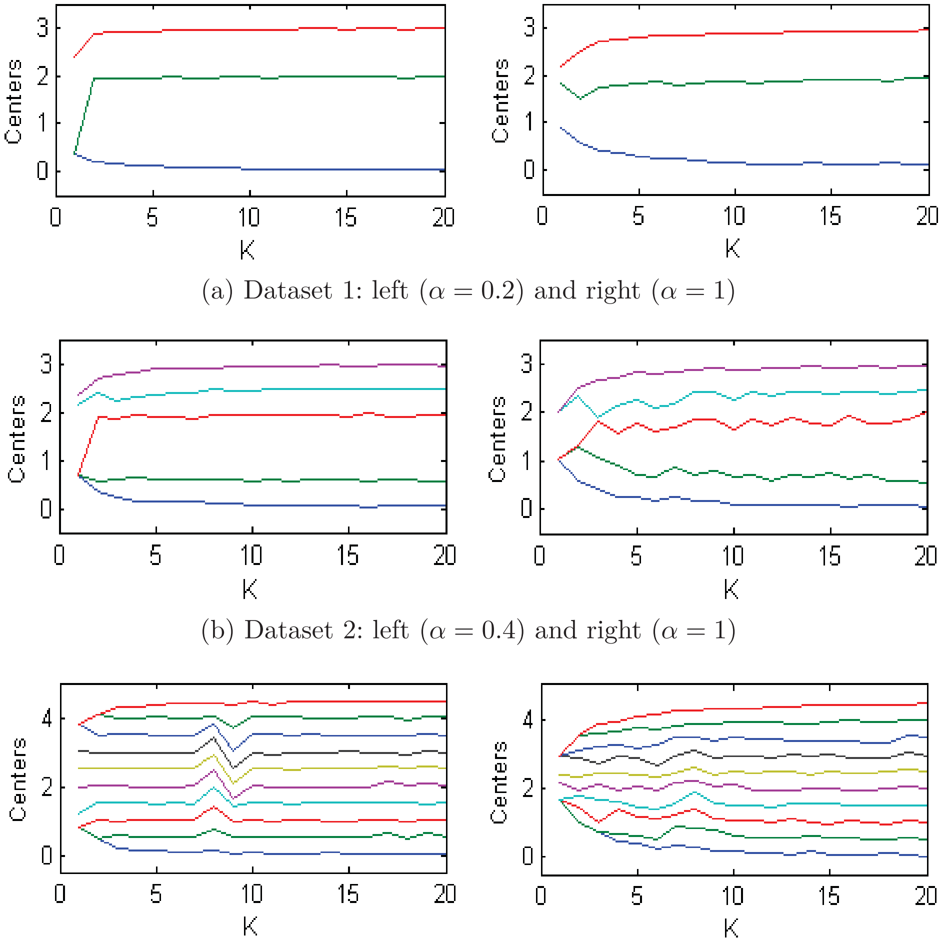


Figure 2.11 – Center positions obtained by our estimation approach in dataset 1 for $K = 1$, $K = 3$ and $K = 6$.

tering process. Different tests are achieved by browsing the whole data points, i.e. $n = N$. Figure 2.13 shows the center positions obtained by our estimation approach on dataset 1 for three different values of $n$. In this figure, we observe that different centers are wrongly located for $n$ equal to 48. However, the centers take their correct locations when $n$ equal to 150 and 1500. Our approach is also tested with sequences of values of $n$ (from 1 to $N$) and the variation of the found centers location is showed in figure 2.14. In several cases, we observe that suitable results can be reached with low sample sizes especially for $\alpha$ between 0.2 and 1, and $K \geq L$. This last inference shows that our approach can be very useful for large scale datasets.

Two other multi-dimensional datasets, Iris data and Yeast data, are tested by the clustering techniques. The datasets, differ in dimension and number of data points and clusters. The datasets were obtained from the machine learning repository at the University of California, Irvine. Iris is four-dimensional dataset which contains 3 classes of 50 points where each class refers to a type of iris plant: iris Versicolor, iris Verginica and iris Setosa. Yeast dataset includes 1484 points characterized by eight features to classify the data points into ten clusters. Our clustering approach was compared with the SRS-$k$-means, the standard $k$-means and the clustering approach based on Bradley refinement of initial centers (called in this paper b-$k$-means) previously described in section 3. The different clustering techniques were used with the same initial centers. The data points are assigned to the cluster having the greatest membership degree. Perfect RSS was used for our approach and the parameter $\alpha$ was fixed to 0.2 for datasets 1 and 3, 0.4 for dataset 2 and 1 for Iris and Yeast datasets.



(a) Dataset 1: left ($\alpha = 0.2$) and right ($\alpha = 1$)

(b) Dataset 2: left ($\alpha = 0.4$) and right ($\alpha = 1$)

(c) Dataset 3: left ($\alpha = 0.2$) and right ($\alpha = 1$)

Figure 2.12 – Variation of the obtained centers for each value of $K$ (from 1 to 20). The sample size $n$ is fixed to the whole dataset size $N$.

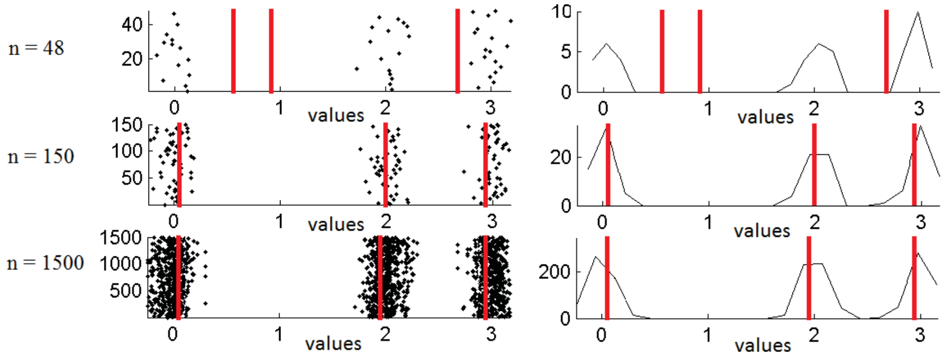


Figure 2.13 – Center positions obtained by our estimation approach in dataset 1 for different sample sizes: $n$=48, $n$=150, and $n$=1500.

The obtained results of the five datasets (dataset 1, dataset 2, dataset 3, Iris and Yeast) were statistically evaluated in terms of classification performance, centers accuracy and clusters validity indices. The first one is commonly expressed as the percentage measure of correctly classified samples which is calculated as follows:

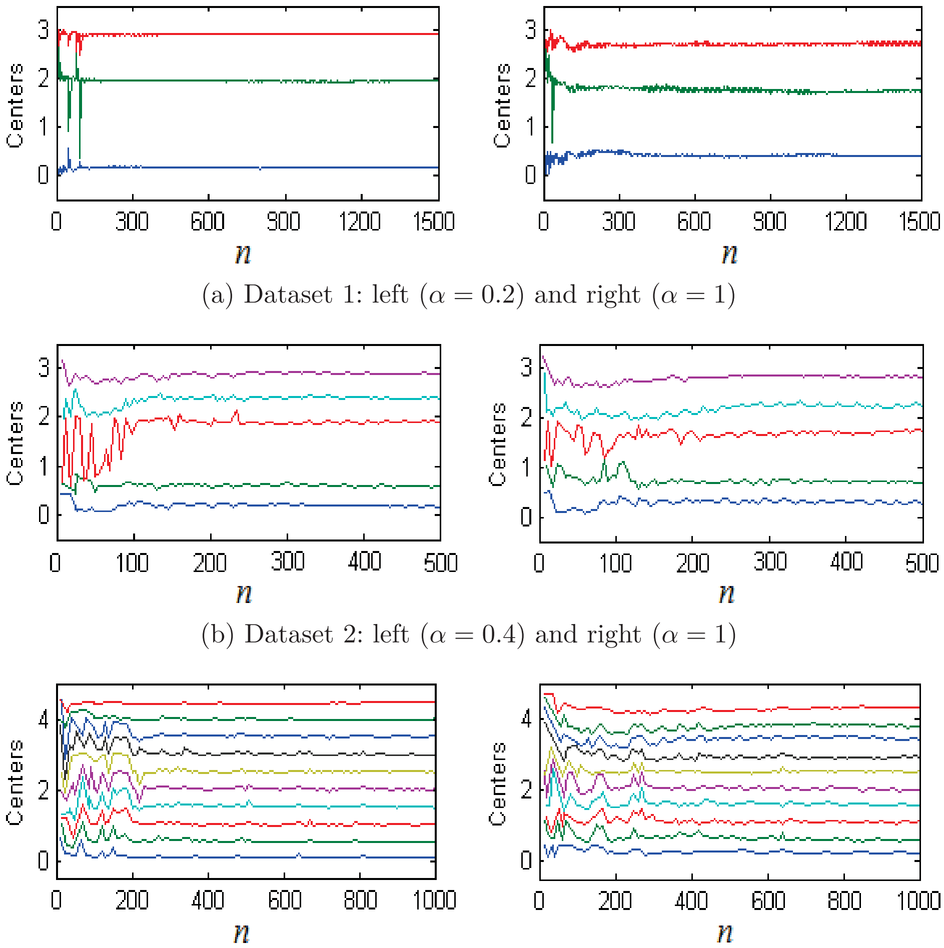
$$\text{Performance} = \frac{N - \text{Misclassified points}}{N} \times 100\% \tag{2.14}$$

where $N$ represents the size of the whole dataset.

The second criteria was used to confirm the relationship deduced in the equation 2.6. A good centers estimation can be obtained when the sum of the variances of center estimators is much less as possible. The center accuracy index is then inversely proportional to the sum of variances and it can be calculated by the following equation:

$$\text{Accuracy} = \frac{1}{\sum_{l=1}^{L} var(\hat{m}_l)} \tag{2.15}$$

The clustering approaches were also compared by three cluster validity measures: Dunn index, Davies-Bouldin index and Silhouette index. Then, the quality of clusters is measured in terms of homogeneity and separation knowing that points within one cluster are similar, while points in different clusters are dissimilar. Dunn index [43] attempts to identify the clusters which are compact and well separated. It is calculated



(a) Dataset 1: left ($\alpha = 0.2$) and right ($\alpha = 1$)

(b) Dataset 2: left ($\alpha = 0.4$) and right ($\alpha = 1$)

(c) Dataset 3: left ($\alpha = 0.2$) and right ($\alpha = 1$)

Figure 2.14 – Variation of the obtained centers for each value of the sample size $n$ (from 1 to $N$). The subpopulation number $K$ is fixed to the center number $L$.

by using the following formula:

$$\text{Dunn index} = \min_{1\leq i\leq L}\left(\min_{1\leq j\leq L}\left(\frac{\delta(C_i,C_j)}{\max_{1\leq k\leq L}\Delta(C_k)}\right)\right) \tag{2.16}$$

where $\delta(C_i, C_j)$ represents the inter-cluster distance between two clusters $C_i$ and $C_j$ which is defined by $\min\{d(x_i, x_j)|x_i \in C_i, x_j \in C_j\}$ and $\Delta(C_k)$ represents the intra-cluster distance of the cluster $C_k$ which is defined by $\max\{d(x_i, x_j)|x_i, x_j \in C_k\}$. $d$ represents the Euclidian distance and $L$ represents the number of clusters. Large values of Dunn index indicate that clusters are compact and well-separated. Davies-Bouldin index [39] is the second cluster validity index which also describes the compactness and the separation of the clusters. It is calculated as follows:

$$\text{Davies-Bouldin index} = \frac{1}{L}\sum_{i=1}^{L}\max_{i\neq j}\left(\frac{\gamma(m_i)+\gamma(m_j)}{\theta(m_i,m_j)}\right) \tag{2.17}$$

where $m_i$ is the center of the $i^{th}$ cluster, $\gamma(m_i)$ is the average distance of all points in the $i^{th}$ cluster to the center $m_i$, $\theta(m_i, m_j)$ is the distance between the two centers $m_i$ and $m_j$, and $L$ is the number of clusters. Davies-Bouldin index has a low score if the clusters are compact and well separated from each other. Therefore, it will has a small value for a good clustering. Silhouette index [96] also describes the compactness and the separation of the clusters. It is calculated as follows:

$$\text{Silhouette index} = \frac{1}{N}\sum_{i=1}^{N}\frac{\beta(x_i)-\alpha(x_i)}{\max\{\alpha(x_i),\beta(x_i)\}} \tag{2.18}$$

where $\alpha(x_i)$ is the average distance of the point $x_i$ to other points in the same cluster, $\beta(x_i)$ is the average distance of the point $x_i$ to the points in its nearest neighbor cluster. Silhouette index represents the average of the ratio in equation 2.18. A larger value indicates a better quality of the clustering result.

Table 2.2, table 2.3 and table 2.4 present the statistics of the different clustering techniques respectively on dataset 1, dataset 2 and dataset 3. In these tables, the different measures present the average values of twenty runs with the same dataset. In the case of ranked-$k$-means, $K$ was used equal to $L$. The parameter $\alpha$ was fixed to 0.2 for dataset 1 and dataset 3, and 0.4 for dataset 2. The Bradley approach, b-$k$-

Table 2.2 – Clustering performances, centers accuracies and cluster validity indices of dataset 1

| Algorithm | Performance | | | Accuracy | | | Cluster validity indices | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Dunn index | | | Davies Bouldin | | | Silhouette | | |
| | $n$=15 | $n$=150 | $n$=1500 | $n$=15 | $n$=150 | $n$=1500 | $n$=15 | $n$=150 | $n$=1500 | $n$=15 | $n$=150 | $n$=1500 | $n$=15 | $n$=150 | $n$=1500 |
| $k$-means | - | - | 84% | - | - | 1.16 | - | - | 0.0014 | - | - | 0.4922 | - | - | 0.8364 |
| b-$k$-means | 70% | 87% | 88% | 243.9 | 1537.6 | 4694.7 | 0.0319 | 0.0319 | 0.0319 | 0.4812 | 0.4889 | 0.4891 | 0.9876 | 0.9876 | 0.9886 |
| SRS-$k$-means | 73% | 81% | 85% | 0.79 | 1.16 | 1.17 | 0.0003 | 0.0014 | 0.0014 | 0.4984 | 0.5252 | 0.4920 | 0.8353 | 0.8364 | 0.8364 |
| Ranked-$k$-means | 92% | 100% | 100% | 131.8 | 1782.3 | 4763.4 | 0.0076 | 0.0379 | 0.0379 | 0.4847 | 0.4859 | 0.4805 | 0.9804 | 0.9833 | 0.9933 |

Table 2.3 – Clustering performances, centers accuracies and cluster validity indices of dataset 2

| Algorithm | Performance | | | Accuracy | | | Cluster validity indices | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Dunn index | | | Davies Bouldin | | | Silhouette | | |
| | $n$=50 | $n$=250 | $n$=500 | $n$=50 | $n$=250 | $n$=500 | $n$=50 | $n$=250 | $n$=500 | $n$=50 | $n$=250 | $n$=500 | $n$=50 | $n$=250 | $n$=500 |
| $k$-means | - | - | 69% | - | - | 0.70 | - | - | 0.0041 | - | - | 0.4228 | - | - | 0.7927 |
| b-$k$-means | 72% | 79% | 80% | 2.03 | 5.75 | 5.07 | 0.0017 | 0.0022 | 0.0022 | 0.4398 | 0.4660 | 0.4692 | 0.7857 | 0.7668 | 0.7711 |
| SRS-$k$-means | 60% | 69% | 70% | 0.79 | 0.78 | 0.76 | 0.0043 | 0.0091 | 0.0091 | 0.4189 | 0.4204 | 0.4150 | 0.7945 | 0.7945 | 0.7935 |
| Ranked-$k$-means | 84% | 92% | 93% | 1.88 | 4.02 | 12.81 | 0.0119 | 0.0119 | 0.0119 | 0.4105 | 0.4061 | 0.4029 | 0.8832 | 0.8615 | 0.8691 |

Table 2.4 – Clustering performances, centers accuracies and cluster validity indices of dataset 3

| Algorithm | Performance | | | Accuracy | | | Cluster validity indices | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Dunn index | | | Davies Bouldin | | | Silhouette | | |
| | $n$=100 | $n$=500 | $n$=1000 | $n$=100 | $n$=500 | $n$=1000 | $n$=100 | $n$=500 | $n$=1000 | $n$=100 | $n$=500 | $n$=1000 | $n$=100 | $n$=500 | $n$=1000 |
| $k$-means | - | - | 57% | - | - | 0.73 | - | - | 0.0072 | - | - | 0.3464 | - | - | 0.8497 |
| b-$k$-means | 69.8% | 80% | 80% | 2.81 | 17.67 | 17.76 | 0.0021 | 0.0027 | 0.0113 | 0.3493 | 0.3485 | 0.3487 | 0.8340 | 0.8466 | 0.8469 |
| SRS-$k$-means | 54% | 57% | 56% | 0.70 | 1.16 | 1.17 | 0.0110 | 0.0027 | 0.0114 | 0.3483 | 0.3423 | 0.3461 | 0.8504 | 0.8488 | 0.8465 |
| Ranked-$k$-means | 75% | 90% | 98% | 2.6 | 8.8 | 28.6 | 0.0278 | 0.0631 | 0.0631 | 0.3333 | 0.3067 | 0.2905 | 0.9161 | 0.9165 | 0.9165 |

Table 2.5 – Clustering performances, centers accuracies and cluster validity indices of Iris dataset

| Algorithm | Performance | | | Accuracy | | | Cluster validity indices | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Dunn index | | | Davies Bouldin | | | Silhouette | | |
| | $n$=30 | $n$=60 | $n$=150 | $n$=30 | $n$=60 | $n$=150 | $n$=30 | $n$=60 | $n$=150 | $n$=30 | $n$=60 | $n$=150 | $n$=30 | $n$=60 | $n$=150 |
| $k$-means | - | - | 59% | - | - | $3.10^{-3}$ | - | - | 0.0438 | - | - | 0.7227 | - | - | 0.6517 |
| b-$k$-means | 61% | 65% | 65% | $2.10^{-3}$ | $2.10^{-3}$ | $2.10^{-3}$ | 0.0890 | 0.0839 | 0.0894 | 0.6632 | 0.6218 | 0.6231 | 0.6827 | 0.6543 | 0.7044 |
| SRS-$k$-means | 64% | 61% | 60% | $2.10^{-3}$ | $3.10^{-3}$ | $3.10^{-3}$ | 0.0370 | 0.0641 | 0.0653 | 0.6546 | 0.6576 | 0.6530 | 0.6563 | 0.6564 | 0.6517 |
| Ranked-$k$-means | 71% | 84% | 92% | $9.10^{-3}$ | $4.10^{-3}$ | $13.10^{-3}$ | 0.0901 | 0.0909 | 0.0909 | 0.7080 | 0.7256 | 0.6883 | 0.7282 | 0.7282 | 0.7282 |

Table 2.6 – Clustering performances, centers accuracies and cluster validity indices of Yeast dataset

| Algorithm | Performance | | | Accuracy | | | Cluster validity indices | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Dunn index | | | Davies Bouldin | | | Silhouette | | |
| | $n$=50 | $n$=500 | $n$=1484 | $n$=50 | $n$=500 | $n$=1484 | $n$=50 | $n$=500 | $n$=1484 | $n$=50 | $n$=500 | $n$=1484 | $n$=50 | $n$=500 | $n$=1484 |
| $k$-means | - | - | 23% | - | - | 13.4 | - | - | 0.0192 | - | - | 1.7272 | - | - | 0.2532 |
| b-$k$-means | 30% | 35% | 35% | 12.02 | 16.51 | 29.01 | 0.0260 | 0.0288 | 0.0227 | 62.6215 | 4.4896 | 4.3254 | 0.1244 | 0.1210 | 0.0542 |
| SRS-$k$-means | 21% | 23% | 23% | 12.0 | 12.2 | 12.6 | 0.0150 | 0.0192 | 0.0190 | 115.4338 | 26.1401 | 1.7239 | 0.0556 | 0.2013 | 0.2511 |
| Ranked-$k$-means | 40% | 49% | 50% | 313 | 475 | 557 | 0.0223 | 0.0298 | 0.0299 | 66.2424 | 3.9218 | 1.3037 | 0.1215 | 0.2803 | 0.2948 |

means, was used with $T = 5$ sub-samples, each sub-sample has a size equal to $n_1$. The $T$ sub-samples constitute the sample to be used for the refinement of initial centers for the clustering process and the sample size $n$ is then equal to $T \times n_1$. For dataset 1, a comparison between ranked-$k$-means, SRS-$k$-means and b-$k$-means are done for three different values of $n$ (15, 150 and 1500 points). For the three values of $n$, the clustering performance of the ranked-$k$-means is superior to the other approaches, the accuracy of the estimated centers is also better and the cluster validity indices show that ranked-$k$-means is also much better then the other approaches in terms of homogeneity into clusters and separation between clusters. Table 2.3 and table 2.4 display also the results for $n$ equal 50, 250 and 500 points and $n$ equal 100, 500 and 1000 points respectively of dataset 2 and dataset 3, and confirms the efficiency of the results of our approach.

Table 2.5 and table 2.6 show also the quantitative performances of the obtained clusters and centers respectively on the Iris dataset and the Yeast dataset. The obtained results show that our approach based on RSS outperforms the standard $k$-means, the b-$k$-means and also the SRS-$k$-means for the different sample sizes. Statistics of the accuracy of the centers $m_{l,RSS}$ show that the ranked-$k$-means is independent to initial centers for several runs of the algorithm. The precision of the centers has an effective impact on the performance of the clustering approaches. The statistics of the clustering performance and the cluster validity indices confirm the efficiency of the ranked-$k$-means. It is notable that the standard fuzzy $k$-means and the SRS-$k$-means produce approximately the same performances and accuracies for $n = N$, i.e. 1500, 500, 1000, 150 and 1484 points respectively of dataset 1, dataset 2, dataset 3, Iris data and Yeast data. It is also remarkable that b-$k$-means outperforms standard $k$-means and SRS-$k$-means for the most indices of performances.

## 5.2 Terahertz datasets

In this section, the different clustering techniques are tested on the THz images of a carbon fiber sample, a fruit grape and a flexure spring. Simultaneous acquisition of visible images, see figure 2.15, allows a comparison between the two types of data. Images in the visible light (first row of figure 2.15) present only a description of per-

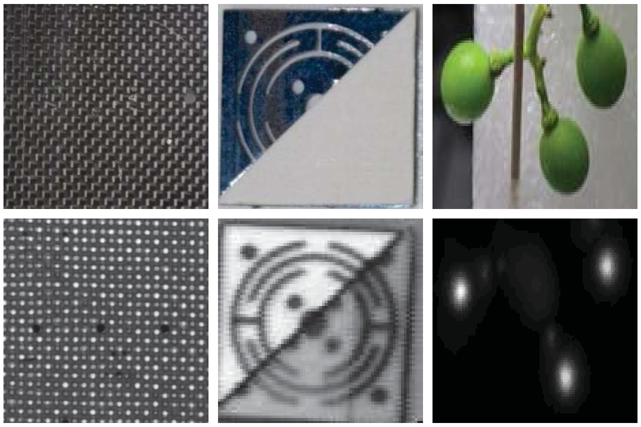


Figure 2.15 – Visible (first row) and THz (second row) images respectively of carbon fiber (first column), flexure spring (second column) and grape (third column). THz images are shown using the maximal amplitude feature.

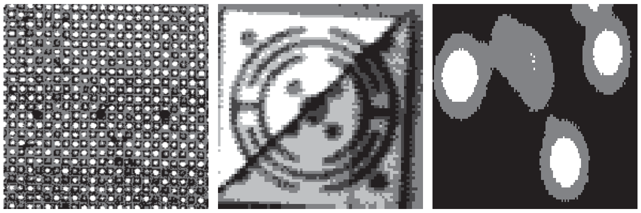


Figure 2.16 – Standard $k$-means segmentation of the three THz images: carbon fiber image (first column), flexure spring image (second column) and grape image (third column). The approach was used respectively with $L$ equal to 3, 3 and 5 clusters for the three images. The AR($p$)/PCA($q$) extracted features were used with $p$=2 and $q$=1 for the carbon and the spring images and $p$=3 and $q$=1 for the grape image.

ceptible objects of the human eye, while in the THz light (second row of figure 2.15), using the standard maximal amplitude (maxA) feature of each pulse (pixel), visible and hidden structures are very well identified in the images. The maxA THz images give us an insight about the structure of the objects and the distribution of the different pixels in the image. For the three maxA THz images, pixels corresponding to large THz reflection are white, while black corresponds to no reflection detected. The segmentation of the carbon fiber image (first column) consists in detecting damaged regions, not visible to the human eye, which can be provided by a hole punch, an end-mill tool or a piece of aluminum scrap. In the second column, the flexure spring

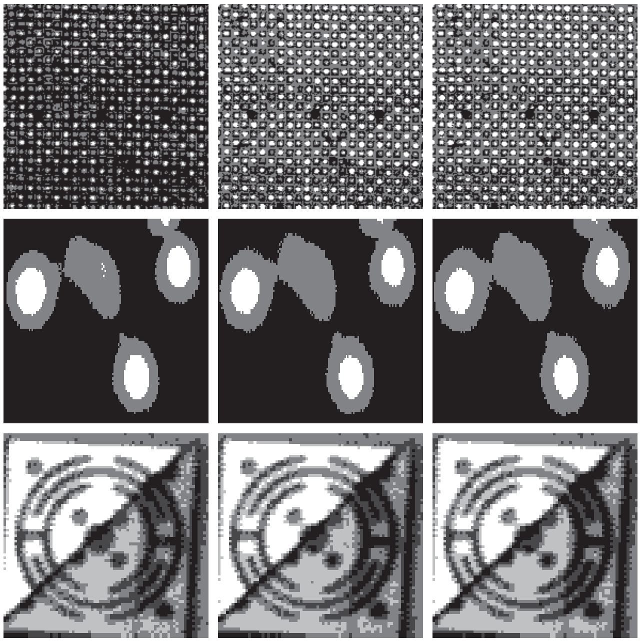


Figure 2.17 – Bradley segmentation of the three THz images: carbon fiber image (first row), grape image (second row) and flexure spring image (third row). The Bradley approach was used respectively with $L = 3$, $L = 3$ and $L = 5$, and the sample size $n$ equal to 30 (first column), 300 (second column) and 3000 (third column). The AR($p$)/PCA($q$) extracted features were used with $p$=2 and $q$=1 for the carbon and the spring images and $p$=3 and $q$=1 for the grape image.

image with one half covered by a business card cutout is shown. In the third column, grape image was shown in both types. The different regions in the THz image represent grape berries, stems, branches and background. Each pixel of the three raw THz images is represented respectively by 572, 700 and 1024 observations in the time domain and its correspondent spectral bands in the frequency domain. We have used in this paper the AR($p$)/PCA($q$) feature extraction method from our previous work in [14], where $p$ and $q$ represent respectively the AR and PCA feature numbers. The method consists to combine the two effective statistical feature extraction models autoregressive (AR) model on the temporal data and principal component analysis (PCA) on the spectral data. The combination AR/PCA has been used to further consolidates the effectiveness of both techniques to extract the pertinent features from the THz images. The pixels of the THz AR($p$)/PCA($q$) data represent, in

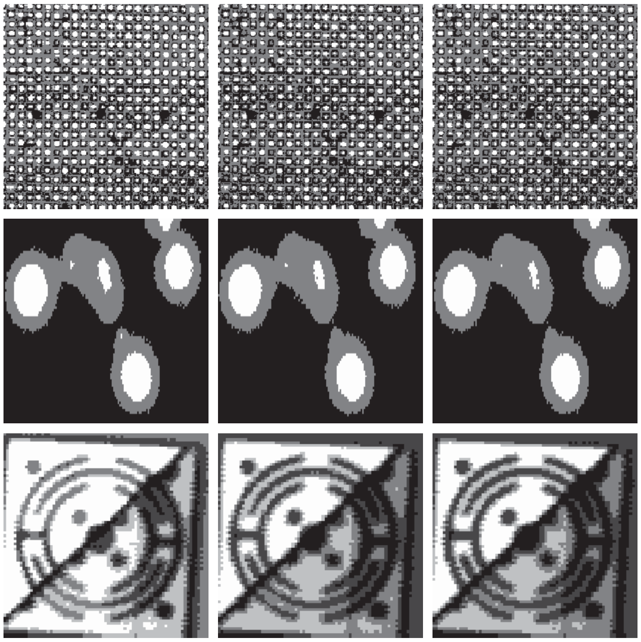


Figure 2.18 – SRS-$k$-means segmentation of THz images of carbon fiber (first row), flexure spring (second row) and grape (third row). For the three images, SRS-$k$-means has used respectively $L = 3$, $L = 3$ and $L = 5$, and the sample size $n$ equal to 30 (first colomn), 300 (second column) and 3000 (third column). The AR($p$)/PCA($q$) extracted features were used with $p$=2 and $q$=1 for the carbon and spring images and $p$=3 and $q$=1 for the grape image.

the RSS, the variable of interest $X$ to be sorted. The different pixels are then sorted according to the concomitant variable $Y$ represented by the standard maxA feature which is highly correlated with the main variable of interest. Ranked-$k$-means was applied with a parameter $\alpha = 1$. The carbon fiber and the spring images were used with $p = 2$ and $q = 1$, while the grape image was used with $p = 3$ and $q = 1$. The different clustering techniques, ranked-$k$-means, SRS-$k$-means, standard $k$-means and b-$k$-means, were applied on the Terahertz images with a parameter $a = 2$, as that was one of the best values found by [24]. The fuzzy segmentation of the three images was employed respectively with 3, 5 and 3 clusters. Figure 2.16 shows the standard $k$-means segmentation of the three THz images. The figure shows that the different regions are wrongly segmented, especially for the carbon fiber and the grape images.

Figures 2.17, 2.18 and 2.19 show the segmentation of the three Terahertz images

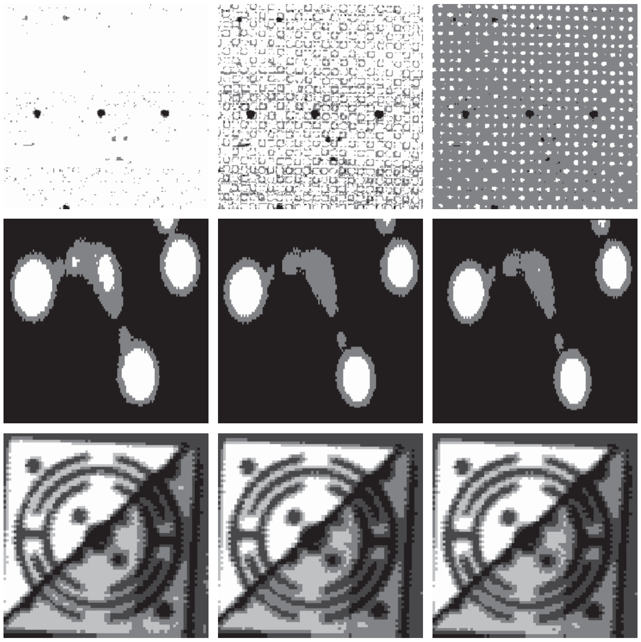


Figure 2.19 – Ranked-$k$-means segmentation of THz images of carbon fiber (first row), flexure spring (second row) and grape (third row). For the three images, ranked-$k$-means has used respectively $L = 3$, $L = 3$ and $L = 5$, and the sample size $n$ equal to 30 (first colomn), 300 (second column) and 3000 (third column). The AR($p$)/PCA($q$) extracted features were used with $p$=2 and $q$=1 for the carbon and spring images and $p$=3 and $q$=1 for the grape image.

respectively for b-$k$-means, SRS-$k$-means and ranked-$k$-means. The segmentation was carried out using three different values of the sample size $n$ (30, 300, and 3000). Figures 2.17 and 2.18 show that the different regions are not well segmented. The obtained regions of the carbon fiber prove that this method gives bad identification of damaged zones for all values of $n$. However in figure 2.19, when increasing the sample size the different regions became well identified by the ranked-$k$-means segmentation and they can be seen as dark gashes along the image. b-$k$-means and SRS-$k$-means applications on the grape image are shown in the second row of figures 2.17 and 2.18. The obtained regions clearly illustrate the limitations of these methods to provide good structure of stems and berries. However, the second row of figure 2.19 presents the regions output of our segmentation approach. White cluster represents the grape berries; gray cluster represents the external layer of grape berries and the stems, and the black cluster represents background and branches. Compared to the b-$k$-means and the SRS-$k$-means segmentations, ranked-$k$-means provides more relevant details of the vine structure, including the stems and the grape berries shape. The third row of figures 2.17, 2.18 and 2.19 show the regions output of b-$k$-means, SRS-$k$-means and ranked-$k$-means clustering using the same AR(2)/PCA(1) data spring. For a small sample size equal to 30, SRS-$k$-means segmentation provides a bad identification of the white cluster, while b-$k$-means provides noisy pixels in the left part of the image. However, increasing the sample size, the obtained regions become more interesting. As well, our approach using the ranked sample, gives promising results for the three values of $n$. Visible and hidden parts of the flexure spring are well segmented especially in the low part of the spring image. The segmentation of the three Terahertz images shows that ranked-$k$-means outperforms the segmentation algorithms standard $k$-means, b-$k$-means and SRS-$k$-means for different values of the sample size.

Visual results of Terahertz images is supported by the statistics shown in table 2.7, table 2.8 and table 2.9. As the ground truth of the THz images is not very precise in our work, we study and show the performances of the clustering techniques only in terms of centers accuracies and cluster validity indices. In the case of ranked-$k$-means, $K$ was used equal to $L$. The parameter $\alpha$ was fixed to 1 for three THz images. A comparison between ranked-$k$-means, SRS-$k$-means and b-$k$-means are done for three different values of $n$ (30, 300 and 3000 pixels). For the three values of $n$, the accuracy of the estimated centers of the ranked-$k$-means is better than the other approaches and the cluster validity indices show that ranked-$k$-means is also much better then the other approaches in terms of homogeneity into clusters and separation between clusters. A good choice of the sample size must achieve a compromise between the segmentation accuracy and the low computational cost.

## 6 Conclusion

A novel clustering approach, called ranked-$k$-means, has been proposed in this paper. Ranked-$k$-means is a two-steps algorithm; the E-step consists to estimate optimal centers by using a new objective function based on ranked set sample, while the C-step consists to classify the observed dataset based on the estimated centers.

Table 2.7 – Centers accuracies and cluster validity indices of the Terahertz carbon fiber image

| Algorithm | Accuracy | | | Cluster validity indices | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Dunn index | | | Davies Bouldin | | | Silhouette | | |
| | $n$=30 | $n$=300 | $n$=3000 | $n$=30 | $n$=300 | $n$=3000 | $n$=30 | $n$=300 | $n$=3000 | $n$=30 | $n$=300 | $n$=3000 |
| $k$-means | - | - | 0.0938 | - | - | 0.0192 | - | - | 0.7563 | - | - | 0.2507 |
| b-$k$-means | 0.1196 | 0.1260 | 0.3678 | 0.0138 | 0.0215 | 0.0252 | 1.0848 | 0.8486 | 0.6327 | 0.5154 | 0.5273 | 0.7331 |
| SRS-$k$-means | 0.0872 | 0.0821 | 0.0912 | 0.0159 | 0.0204 | 0.0199 | 1.2781 | 1.1203 | 0.9721 | 0.5525 | 0.5512 | 0.7222 |
| Ranked-$k$-means | 1.8201 | 2.5201 | 3.7132 | 0.0218 | 0.0248 | 0.0297 | 1.0981 | 0.8951 | 0.7452 | 0.5214 | 0.5320 | 0.7012 |

Table 2.8 – Centers accuracies and cluster validity indices of the Terahertz grape image

| Algorithm | Accuracy | | | Cluster validity indices | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Dunn index | | | Davies Bouldin | | | Silhouette | | |
| | $n$=30 | $n$=300 | $n$=3000 | $n$=30 | $n$=300 | $n$=3000 | $n$=30 | $n$=300 | $n$=3000 | $n$=30 | $n$=300 | $n$=3000 |
| $k$-means | - | - | 0.2291 | - | - | 0.0129 | - | - | 0.5763 | - | - | 0.7354 |
| b-$k$-means | 0.1944 | 0.2090 | 0.2325 | 0.0016 | 0.0022 | 0.0035 | 0.8043 | 0.6298 | 0.4868 | 0.7950 | 0.7993 | 0.8212 |
| SRS-$k$-means | 0.1039 | 0.1920 | 0.1911 | 0.0027 | 0.0031 | 0.0031 | 0.5743 | 0.4718 | 0.4842 | 0.8150 | 0.8031 | 0.8214 |
| Ranked-$k$-means | 2.7122 | 5.8139 | 6.9863 | 0.0058 | 0.0068 | 0.0127 | 0.5504 | 0.5201 | 0.5201 | 0.8019 | 0.8103 | 0.8230 |

Table 2.9 – Centers accuracies and cluster validity indices of the Terahertz flexure spring image

| Algorithm | Accuracy | | | Cluster validity indices | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Dunn index | | | Davies Bouldin | | | Silhouette | | |
| | $n$=30 | $n$=300 | $n$=3000 | $n$=30 | $n$=300 | $n$=3000 | $n$=30 | $n$=300 | $n$=3000 | $n$=30 | $n$=300 | $n$=3000 |
| $k$-means | - | - | 0.2941 | - | - | 0.0063 | - | - | 0.8389 | - | - | 0.5607 |
| b-$k$-means | 0.3281 | 0.3299 | 0.3300 | 0.0038 | 0.0054 | 0.0069 | 0.9673 | 0.8489 | 0.8018 | 0.5409 | 0.5542 | 0.5579 |
| SRS-$k$-means | 0.0517 | 0.2135 | 0.2931 | 0.0051 | 0.0051 | 0.0054 | 0.8074 | 0.8074 | 0.8074 | 0.5311 | 0.5311 | 0.5311 |
| Ranked-$k$-means | 0.8969 | 3.1852 | 4.9820 | 0.0052 | 0.0055 | 0.0059 | 0.8619 | 0.8489 | 0.8018 | 0.5409 | 0.5519 | 0.5519 |

Ranked-$k$-means is essentially less sensitive to the initialization of the centers. The performance of ranked-$k$-means is valorized regarding to standard $k$-means, $k$-means based on SRS sample and $k$-means based on Bradley refinement of initial centers on several datasets, specially in the case of the Terahertz imaging.

It is shown that suitable results can be reached with a low sample size $n$ especially for $\alpha$ about between 0.2 and 1, and a subpopulation number $K$ superior or equal to the cluster number $L$. This last inference shows that our approach can be very useful for large-scale datasets. Note that, precise values of $\alpha$, $K$ and $n$, which are important to guarantee good clustering results, haven't been addressed in this paper.

In further work, we will deal with the feature selection and the estimation of the parameters $\alpha$, $K$, and $n$. For instance, the parameter $\alpha$ can be estimated by a Bayesian interpretation to equation (2.8). Then, a prior probability to $\alpha$ can be added in our work to estimate its optimal value. Similar steps to the work of Allili and Ziou [8] can be an issue to develop this idea.

## Acknowledgment

The authors would like to thank the Professor John F. Federici of New Jersey Institute of Technology USA and Professor Thomas Tongue of Zomega Terahertz Corporation for the Terahertz measurements.

# Chapitre 3

# Segmentation d'images THz utilisant $K$-means basée sur la pondération d'attributs et l'échantillonnage aléatoire

Dans le chapitre précèdent, nous avons proposé une approche de segmentation d'images Térahertz basée sur le $K$-means et l'échantillonnage ordonné. Cette approche est moins sensible aux conditions de départ, toutefois, elle est face à des défis pour la sélection des caractéristiques pertinentes et le choix de la caractéristique concomitante utilisée pour trier les pixels. Le deuxième chapitre aborde la sélection des données à travers la pondération de caractéristiques et l'échantillonnage aléatoire simple pour la classification des pixels en vue d'une segmentation des images Térahertz. Une estimation automatique de la taille de l'échantillon aléatoire et du nombre de caractéristiques sélectionnées sont également proposés.

Dans ce chapitre, nous présentons un article intitulé **Terahertz image segmentation using $K$-means clustering based on weighted feature learning and random pixel sampling** publié dans le journal international de Elsevier **Neurocomputing**, 2016 [13]. J'ai réalisé, validé et rédigé ce travail sous la supervision du professeur Djemel Ziou. Une version compacte de ce travail a été publiée dans la conférence internationale **IEEE Computer Vision and Pattern Recognition workshops (CVPR2015)**, Boston, États-Unis, 2015, intitulée **Automated feature weighting and random pixel sampling in $K$-means clustering for Terahertz image segmentation** [16].

# Terahertz image segmentation using $k$-means clustering based on weighted feature learning and random pixel sampling

Mohamed Walid Ayech
Département d'informatique, Université de Sherbrooke,
Sherbrooke, Québec, Canada J1K 2R1
walid.ayech@usherbrooke.ca

Djemel Ziou
Département d'informatique, Université de Sherbrooke,
Sherbrooke, Québec, Canada J1K 2R1
djemel.ziou@usherbrooke.ca

**Keywords**: Segmentation, Terahertz imaging, $k$-means, simple random sampling, feature weighting, dispersion.

### Abstract

Terahertz (THz) imaging is an innovative technology of imaging which can supply a large amount of data unavailable through other sensors. However, the higher dimension of THz images can be a hurdle to their display, their analysis and their interpretation. In this study, we propose a weighted feature space and a simple random sampling in $k$-means clustering for THz image segmentation. Our approach consists to estimate the expected centers, select the relevant features and their scores, and classify the observed pixels of THz images. Automatic estimation of the random sample size and the selected feature number are also proposed in this paper. Our approach is more appropriate for achieving the best compactness inside clusters, the best discrimination of features, and the best tradeoff between the clustering accuracy and the low computational cost. Our approach of segmentation is evaluated by measuring performances and appraised by a comparison with some related works.

# 1 Introduction

Terahertz radiation (T-ray) refers to the region of electromagnetic spectrum occupying the band of frequencies from 0.1 to 10 THz and bounded by microwave and infrared bands. Compared to X-ray, infrared, visible and microwave, Terahertz image automatic analysis and interpretation are in its infancy. However, the advances in THz acquisition technologies open the door to practical use in several areas such as remote sensing, medical diagnosis, and security. The reader can find more about applications in [63, 54, 82, 83, 53, 69, 70, 26, 71, 65, 112]. Terahertz radiations are non invasive and penetrate dry and non-metallic objects such as paper, wood, cloth, etc.

Terahertz images can be acquired by acquisition in both active or passive modes. Terahertz imaging in the active mode is formed by measures of sequences of chronological series or signals reflected from or transmitted through a sample. Each series can be represented by several bands or features (e.g. 1024 features) which characterise one pixel and the combination of these series into rows and columns constitutes the raw Terahertz data cube (e.g. the $R \times C \times P$ cube in figure 3.1, where $R$, $C$ and $P$ represent respectively the number of rows, columns and features). Beyond the acquisition, the Terahertz image segmentation has been studied in [50, 101, 110, 109, 113]. The $k$-means based clustering is the most popular used technique [14, 15, 77, 24, 68]. However, the high dimensionality of THz images lead to some new challenges for relevant feature selection. Indeed, the relevant features can be embedded only on few bands [54, 22]. For this reason, in several related works, some measures from the whole time series are used, such as the amplitude of the maximal pick and the time delay of the maximal pick of the time series [54, 50, 22]. These measures remain insufficient to characterize the different objects of Terahertz images [14]. However, existing $k$-means algorithms deal all features with equal weights. Moreover, making use of the whole data into the clustering process decreases the $k$-means performances. It is thus necessary to integrate *data sampling* designs and *feature weighting* methods into $k$-means algorithm to extract relevant bands from the vast Terahertz data observations. Data sampling and feature weighting techniques can improve the efficiency and the accuracy of the analysis.

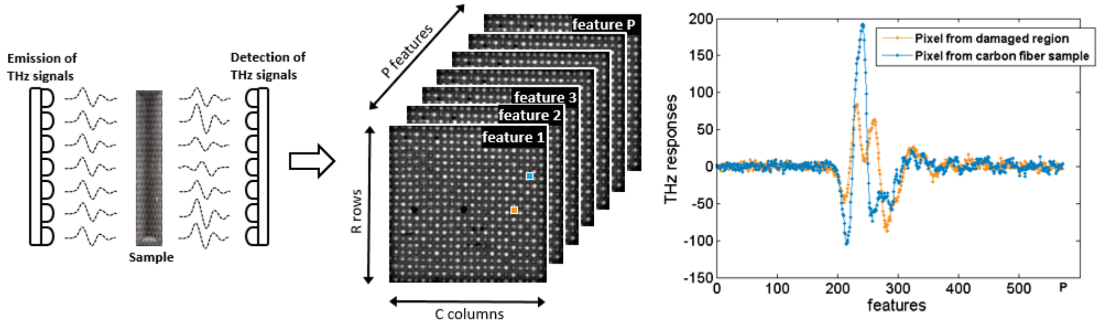


Figure 3.1 – Schematic of Terahertz imaging formation in transmission mode. (a) shows an interaction of T-rays with a sample of carbon fiber. The THz signals are projected to the sample, interacted with it and then detected to constitute the THz data cube in (b). The detected T-rays form the different regions in the sample. The obtained THz data in (b) is represented by $R \times C$ pixels and characterized by $P$ raw features. Two pixels are colored in blue and orange which belong respectively to a typical region of the carbon fiber and a damaged zone. (c) contains two different THz responses colored in blue and orange which correspond respectively to the two pixels in the Terahertz image in (b) with the same color.

In this paper we propose the use of simple random sampling (SRS) and feature weighting. SRS sampling has been used in the works of Ayech and Ziou [15]. They have proposed an approach called SRS-$k$-means which consists to use the $k$-means technique under the SRS sample to avoid the use of the whole set of pixels. In this paper, we propose to reformulate the SRS-$k$-means by selecting the relevant features using a weighting strategy. The weighting methods have been used in [64, 86, 9] to analyze heart diseases data, Australian credit card data and other datasets from the UCI Machine Learning Repository. They have proposed to assign weights to features iteratively updated during the clustering process. Our main contributions consist to integrate the feature weighting and the SRS sampling into the $k$-means clustering. Our approach is called pixel sampling and feature selection in $k$-means clustering (SS-$k$-means) and consists to learn the weights of features according to their clustering importance and introduce the simple random sampling design into the $k$-means process. Note that the SS-$k$-means would be more appropriate for achieving the best compactness inside clusters, the best discrimination of features, and the best tradeoff between the clustering accuracy and the low computational cost. There are four main differences with the state of art: 1) the high dimensional data which are the THz images (1024 bands) and the feature space used; 2) the feature weighting formulation; 3) the combination of the feature weighting and the SRS sampling; 4) the automatic estimation of the sample size and the selected feature number.

The paper is organized as follows: in Section 2, we present an insight about related works of various THz imaging applications. Moreover, a background of $k$-means clustering algorithms has been detailed. We show its limitations and propose subsequently, in Section 3, a novel approach to overcome these limitations. Our approach of segmentation is compared to $k$-means, SRS-$k$-means, KHM, GMM and W-$k$-means on Terahertz images. The results are illustrated and discussed in Section 4.

# 2 Background

## 2.1 Terahertz imaging

Terahertz images are formed by capturing T-rays reflected from or transmitted through objects. Water and moisture objects highly absorb the THz radiations, however, dry objects (such as paper, cloth, wood, and plastic) are transparent to THz radiations and provide no significant reflected radiations. Metals are opaque to T-rays and reflect most incoming radiations. Other interesting materials, which offer specific THz radiations, are detailed in [54, 21]. Terahertz image is formed by several bands (e.g. 1024 bands). The high dimensionality of Terahertz images leads to some new challenges for relevant feature selection. The features are used for the segmentation of Terahertz images. In the most related works, classification of features is used for Terahertz image segmentation.

Numerous works have been proposed to segment Terahertz images. Some works are summarized in this section in terms of feature space used and classification or clustering techniques. The basic feature space is the raw time series of THz images [21]. Other feature spaces are obtained by using Fourier or Wavelet transforms [101, 110, 109]. The feature space can be used with only one band or with several bands. In the first case, the choice of the band can be priori fixed either from the time or the spectral spaces of the THz image [50]. Some measures from the shape of the entire time series or other spectral transform are used, such as the amplitude and the time delay of the maximal pick of time series [50, 22]. In the second case, several bands are used, such as the full time series of THz image, the full spectral amplitude, and a collection of some bands from time series [50, 110, 22, 21]. Some authors are proposed to reduce the feature space by using AR, ARMA, PCA, and decision tree [50, 109, 113, 14, 31, 85]. THz image segmentation is generally performed in terms of classification, such as Mahalanobis classifier, SVM and neural networks [50, 110, 109, 113], and clustering, such as $k$-means, ISODATA, hierarchical chameleon, and KHM [101, 14, 15, 22, 21, 31, 85].

In the most related works, $k$-means clustering has been shown efficient for the Terahertz image segmentation [15, 22, 21, 31, 85]. However, $k$-means techniques deal all the features with equal weights. Besides, making use of the whole data into the clustering process may decrease the $k$-means performances. We show in this paper the effect of both random sampling of the pixels and weighting method of the features in the $k$-means clustering accuracy. Our approach is compared with some related works presented in section 2.2.

## 2.2 Clustering based $k$-means algorithms

Let $X = \{x_1, \cdots, x_N\}$ constitutes a set of $N$ data points. Each data point $x_j = (x_{j1}, x_{j2}, \cdots, x_{jP})$ represents a feature vector in $P$-dimensional space. In this paper, we are interested in classifying the dataset $X$ into homogeneous clusters. One of the well known clustering techniques, $k$-means [68, 77, 24], is largely used to regroup datasets into a pre-defined number $L$ of clusters. The $k$-means type algorithms consist to assign each data point to the cluster having the closest center using either hard or soft decision rule [77, 24]. Tables 3.1 and 3.2 present a summary of some related works on $k$-means algorithms where each one is characterized by the size of the sample used, the objective function formula, the membership degrees $u_{jl}$ of data points to clusters, the centers $m_{lp}$ of clusters, and the weights $w_p$ of features (or $w_{pl}$ of feature/cluster relationships). From this table, we deduce three important remarks. The first one concerns the sample size used into the clustering techniques. For $k$-means, KHM ($k$-

Table 3.1 – Summary of some related works on $k$-means clustering techniques in terms of sample size and objective function. $d(x_{jp}, m_{lp})$ is the Euclidian distance between $x_{jp}$ and $m_{lp}$. $u_{jl}$ is the membership degree of the $j^{th}$ data point to the $l^{th}$ cluster. $m_{lp}$ is the $l^{th}$ center for the $p^{th}$ feature. $w_p$ is the $p^{th}$ feature weight.

| Methods | Sample size | Objective functions |
|---|---|---|
| $k$-means [24] | $N$ | $\sum_{p=1}^{P}\sum_{l=1}^{L}\sum_{j=1}^{N} u_{jl}^{a} d(x_{jp}, m_{lp})$ |
| KHM [111] | $N$ | $\sum_{p=1}^{P}\sum_{j=1}^{N} \frac{L}{\sum_{l=1}^{L} \frac{1}{d(x_{jp}, m_{lp})^{q}}}$ |
| SRS-$k$-means [15] | $n$ (Given) | $\sum_{p=1}^{P}\sum_{l=1}^{L}\sum_{j=1}^{n} u_{jl}^{a} d(x_{jp}, m_{lp})$ |
| Ranked-$k$-means [15] | $n$ (Given) | $\sum_{p=1}^{P}\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{r=1}^{n/K} u_{rkl}^{a} \times (d(x_{rkp}, m_{lp}) + \alpha d(v_{kp}, m_{lp}))$ |
| EW-$k$-means [75] | $N$ | $\sum_{p=1}^{P}\sum_{j=1}^{N} u_{jl} w_{pl} d(x_{jp}, m_{lp}) + \delta w_{pl} log(w_{pl})$ |
| SCAD [56] | $N$ | $\sum_{p=1}^{P}\sum_{j=1}^{N} u_{jl}^{a} w_{pl} d(x_{jp}, m_{lp}) + \delta_l w_{pl}^{2}$ |
| HW-$k$-means [64] | $N$ | $\sum_{p=1}^{P}\sum_{l=1}^{L}\sum_{j=1}^{N} u_{jl} w_{p}^{b} d(x_{jp}, m_{lp})$ |
| SW-$k$-means [86] | $N$ | $\sum_{p=1}^{P}\sum_{l=1}^{L}\sum_{j=1}^{N} u_{jl}^{a} w_{p}^{b} d(x_{jp}, m_{lp})$ |
| MW-$k$-means [9] | $N$ | $\sum_{p=1}^{P}\sum_{l=1}^{L}\sum_{j=1}^{N} u_{jl} w_{p}^{b} \lvert x_{jp} - m_{lp}\rvert^{b}$ |
| Our approach | $n$ (Computed) | $\sum_{p=1}^{P} w_{p}^{b} \frac{\sum_{l=1}^{L}\sum_{j=1}^{n} u_{jl}^{a} d(x_{jp}, m_{lp})}{(\sum_{j=1}^{n} d(x_{jp}, m_{p}))^{c}}$ |

harmonic means), EW-$k$-means (entropy weighting $k$-means), SCAD (simultaneous clustering and attribute discrimination), HW-$k$-means (hard weighting $k$-means), SW-$k$-means (soft weighting $k$-means) and MW-$k$-means (Minkowsky weighting $k$-means) techniques, the whole population ($N$ points) is used into the clustering process, while only a small sample of $n$ points ($n \ll N$) from the observed population is used into SRS-$k$-means and ranked-$k$-means techniques. The choice of the sample size $n$ is given by the authors. The second remark concerns the objective function. The standard $k$-means objective function consists to minimize the arithmetic mean of distances from data points to cluster centers. The KHM is an extended approach of $k$-means where its objective function consists to minimize the harmonic mean of distances from data points to cluster centers. The SRS-$k$-means objective function is the same of the standard $k$-means using $n$ (the sample size) instead of $N$ (the whole population size). In the case of ranked-$k$-means, the objective function incorporates a rank constraint and aims at keeping the closeness of the cluster centers around the nearest subpopulation means computed from the ranked set sample. The parameter $\alpha$ in the second term controls the effect of the order statistic of data points. EW-$k$-means objective function is constituted by two terms, the first one is similar to the standard

Table 3.2 – Summary of some related works on $k$-means clustering techniques in terms of membership degrees, centers and feature weights. $d(x_{jp}, m_{lp})$ is the Euclidian distance between $x_{jp}$ and $m_{lp}$. $u_{jl}$ is the membership degree of the $j^{th}$ data point to the $l^{th}$ cluster. $m_{lp}$ is the $l^{th}$ center for the $p^{th}$ feature. $w_p$ is the $p^{th}$ feature weight.

| Methods | Membership degrees | Centers | Feature weights |
|---|---|---|---|
| $k$-means [24] | $\left(\sum_{h=1}^{L}\left(\frac{\sum_{p=1}^{P} d(x_{jp},m_{lp})}{\sum_{p=1}^{P} d(x_{jp},m_{hp})}\right)^{\frac{1}{a-1}}\right)^{-1}$ | $\frac{\sum_{j=1}^{N} u_{jl}^{a} x_{jp}}{\sum_{j=1}^{N} u_{jl}^{a}}$ | $\frac{1}{P}$ |
| KHM [111] | $\frac{\sum_{p=1}^{P} d(x_{jp},m_{lp})^{-q-2}}{\sum_{h=1}^{L}\sum_{p=1}^{P} d(x_{jp},m_{hp})^{-q-2}}$ | $\frac{\sum_{j=1}^{N} u_{jl} v_j x_{jp}}{\sum_{j=1}^{N} u_{jl} v_j}$ | $\frac{1}{P}$ |
| SRS-$k$-means [15] | $\left(\sum_{h=1}^{L}\left(\frac{\sum_{p=1}^{P} d(x_{jp},m_{lp})}{\sum_{p=1}^{P} d(x_{jp},m_{hp})}\right)^{\frac{1}{a-1}}\right)^{-1}$ | $\frac{\sum_{j=1}^{n} u_{jl}^{a} x_{jp}}{\sum_{j=1}^{n} u_{jl}^{a}}$ | $\frac{1}{P}$ |
| Ranked-$k$-means [15] | $\left(\sum_{h=1}^{L}\left(\frac{\sum_{p=1}^{P}(d(x_{rkp},m_{lp})+\alpha d(v_{kp},m_{lp}))}{\sum_{p=1}^{P}(d(x_{rkp},m_{lp})+\alpha d(v_{kp},m_{lp}))}\right)^{\frac{1}{a-1}}\right)^{-1}$ | $\frac{\sum_{k=1}^{K}\sum_{r=1}^{n/K} u_{rkl}^{a}(x_{rkp}+\alpha v_{kp})}{(\alpha+1)\sum_{k=1}^{K}\sum_{r=1}^{n/K} u_{rkl}^{a}}$ | $\frac{1}{P}$ |
| EW-$k$-means [75] | 1, if $l = \arg\min_l \sum_{p=1}^{P} w_{pl} d(x_{jp}, m_{lp})$ 0, if $t \neq l$ | $\frac{\sum_{j=1}^{N} u_{jl} x_{jp}}{\sum_{j=1}^{N} u_{jl}}$ | $\frac{exp-\frac{1}{\delta}\sum_{j=1}^{N} u_{jl} d(x_{jp},m_{lp})}{\sum_{t=1}^{P} exp-\frac{1}{\delta}\sum_{j=1}^{N} u_{jl} d(x_{jt},m_{lt})}$ |
| SCAD [56] | $\left(\sum_{h=1}^{L}\left(\frac{\sum_{p=1}^{P} w_{pl} d(x_{jp},m_{lp})}{\sum_{p=1}^{P} w_{pl} d(x_{jp},m_{hp})}\right)^{\frac{1}{a-1}}\right)^{-1}$ | $\frac{\sum_{j=1}^{N} u_{jl}^{a} x_{jp}}{\sum_{j=1}^{N} u_{jl}^{a}}$ | $\frac{1}{n} + \frac{1}{2\delta_l}\sum_{j=1}^{N} u_{jl}^{a}(\frac{1}{P} d(x_j - m_l) - d(x_{jp} - m_{lp}))$ |
| HW-$k$-means [64] | 1, if $l = \arg\min_l \sum_{p=1}^{P} w_p^b d(x_{jp}, m_{lp})$ 0, if $t \neq l$ | $\frac{\sum_{j=1}^{N} u_{jl} x_{jp}}{\sum_{j=1}^{N} u_{jl}}$ | $\frac{(1/(\sum_{l=1}^{L}\sum_{j=1}^{N} u_{jl} d(x_{jp},m_{lp})))^{1/(b-1)}}{\sum_{t=1}^{P}(1/(\sum_{l=1}^{L}\sum_{j=1}^{N} u_{jl} d(x_{jt},m_{lt})))^{1/(b-1)}}$ |
| SW-$k$-means [86] | $\left(\sum_{h=1}^{L}\left(\frac{\sum_{p=1}^{P} w_p^b d(x_{jp},m_{lp})}{\sum_{p=1}^{P} w_p^b d(x_{jp},m_{hp})}\right)^{\frac{1}{a-1}}\right)^{-1}$ | $\frac{\sum_{j=1}^{N} u_{jl}^{a} x_{jp}}{\sum_{j=1}^{N} u_{jl}^{a}}$ | $\frac{(1/(\sum_{l=1}^{L}\sum_{j=1}^{N} u_{jl}^{a} d(x_{jp},m_{lp})))^{1/(b-1)}}{\sum_{t=1}^{P}(1/(\sum_{l=1}^{L}\sum_{j=1}^{N} u_{jl}^{a} d(x_{jt},m_{lt})))^{1/(b-1)}}$ |
| MW-$k$-means [9] | 1, if $l = \arg\min_l \sum_{p=1}^{P} w_p^b \lvert x_{jp} - m_{lp}\rvert^b$ 0, if $t \neq l$ | $\frac{\sum_{j=1}^{N} u_{jl} x_{jp}}{\sum_{j=1}^{N} u_{jl}}$ | $\frac{(1/(\sum_{l=1}^{L}\sum_{j=1}^{N} u_{jl}\lvert x_{jp}-m_{lp}\rvert))^{1/(b-1)}}{\sum_{t=1}^{P}(1/(\sum_{l=1}^{L}\sum_{j=1}^{N} u_{jl}\lvert x_{jt}-m_{lt}\rvert))^{1/(b-1)}}$ |
| Our approach | $\left(\sum_{h=1}^{L}\left(\frac{\sum_{p=1}^{P} w_p^b \frac{d(x_{jp},m_{lp})}{(\sum_{j=1}^{n} d(x_{jp},m_p))^c}}{\sum_{p=1}^{P} w_p^b \frac{d(x_{jp},m_{hp})}{(\sum_{j=1}^{n} d(x_{jp},m_p))^c}}\right)^{\frac{1}{a-1}}\right)^{-1}$ | $\frac{\sum_{j=1}^{n} u_{jl}^{a} x_{jp}}{\sum_{j=1}^{n} u_{jl}^{a}}$ | $\frac{\left(\frac{(\sum_{j=1}^{n} d(x_{jp},m_p))^c}{\sum_{l=1}^{L}\sum_{j=1}^{n} u_{jl}^{a} d(x_{jp},m_{lp})}\right)^{1/(b-1)}}{\sum_{t=1}^{P}\left(\frac{(\sum_{j=1}^{n} d(x_{jt},m_t))^c}{\sum_{l=1}^{L}\sum_{j=1}^{n} u_{jl}^{a} d(x_{jt},m_{lt})}\right)^{1/(b-1)}}$ |

$k$-means and the second one correspond to the negative weight entropy. The parameter $\delta$ controls the effect of the second term. The SCAD objective function has also two terms. The first one is also similar to the standard $k$-means, while the second one is the squared feature/cluster weights. The minimization of this term is achieved to promote the effect of the feature/cluster weights into the clustering process, specially when the features are equally weighted. The parameter $\delta_l$ controls the effect of the second term. HW-$k$-means, SW-$k$-means and MW-$k$-means objective functions correspond to a weighted sum of within-cluster dispersions. HW-$k$-means and SW-$k$-means algorithms use the Euclidian distance, while the MW-$k$-means algorithm uses the Minkowsky distance. The third remark concerns the parameters estimated from these objective functions. The $k$-means, KHM, SRS-$k$-means and ranked-$k$-means algorithms deal all the features with equal weights. While, EW-$k$-means, SCAD, HW-$k$-means, SW-$k$-means and MW-$k$-means affect to each feature (or feature/cluster relationship) a weight iteratively adjusted during the clustering process. We note that standard $k$-means, SW-$k$-means and SRS-$k$-means are particular cases of the SS-$k$-means. Indeed, the SRS-$k$-means can be obtained from SS-$k$-means when $b = c = 0$, the W-$k$-means when $n = N$ and $c = 0$, and the $k$-means when $n = N$ and $b = c = 0$. There are four main differences of our approach with the state of art: 1) a small random sample is used into the clustering. Its size $n$ is automatically estimated and accurately representative of the whole population; 2) the objective function use a new term $\sum_{j=1}^{n} d(x_{jp}, m_p)$ incorporating the effect of the dispersion of the global data into the clustering; 3) the feature weights and the membership degrees depend on the global data dispersions; 4) the combination of the feature weighting and the SRS sampling. For the rest of our paper, SW-$k$-means approach will be denoted W-$k$-means for more writing simplicity. In the following subsections, we detail the SRS-$k$-means and the W-$k$-means algorithms which constitute the baseline of our approach.

### SRS-$k$-means algorithm

SRS-$k$-means [15] consists to combine the simple random sampling (SRS) [38, 104] and the $k$-means clustering technique [77, 24, 68]. A representative sample $X_{SRS}$ of $n$ points from the observed population $X$ of $N$ points is then randomly selected and regrouped into homogeneous clusters in order to get conclusions about the centers.

The main steps necessary to select a SRS data are summarized as follows:

```
$SRS(X, n)$ algorithm
1. Develop a population list of all the elements of the studied population and assign each element a number to be able to access to the population.
2. Generate a list of $n$ random numbers.
3. Select the elements $\{x_1, ..., x_n\}$ that have numbers corresponding to the generated random number list and save them in a dataset denoted $X_{SRS}$.
4. Return $X_{SRS}$.
```

Figure 3.2 shows an example of a dataset $X$ (on the left) and a small sample $X_{SRS}$ (on the right) distributed into two clusters. We show that only 1% of the population can represent the data under study and allows providing inferential statistics for the whole data clustering. SRS-$k$-means is a two-step algorithm. The E-step (estimation step) consists to classify the $X_{SRS}$ into $L$ clusters; each one is represented by its own estimated center $m_l$. The C-step (clustering step) consists then to affect each point from the observed data $X$ to the nearest cluster represented by its estimated center.

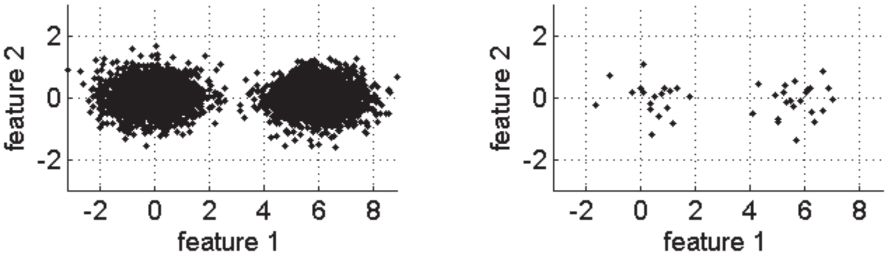
The SRS-$k$-means technique appears faster than the traditional $k$-means; however, all the features are dealt with equal importance into the clustering process in spite of some features can be noisy or uninformative.

### W-$k$-means algorithm

W-$k$-means [86] consists to assign weights to features into $k$-means process. Let $X$ be a set of $N$ data points where each one represents a feature vector in $P$-dimensional space and $\mathbf{w}$ is a vector of $P$ feature weights. W-$k$-means consists to minimize a weighted sum of within-cluster dispersions, reformulated as follows:

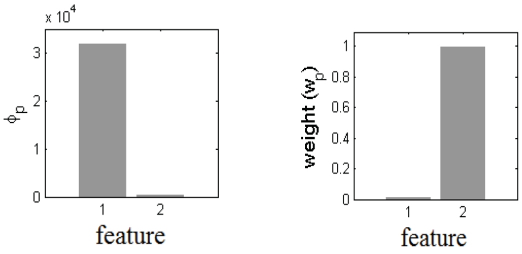
$$J(\Phi, \mathbf{w}) = \sum_{p=1}^{P} w_p^b \varphi_p, \tag{3.1}$$

where $\Phi = (\varphi_1, \cdots, \varphi_P)$ and $\mathbf{w} = (w_1, \cdots, w_P)$ represent two vectors of $P$ variables, $\varphi_p = \sum_{l=1}^{L} \sum_{j=1}^{N} u_{jl}^a d(x_{jp}, m_{lp})$ correspond to the within-cluster dispersion of the $p^{th}$ feature, $w_p$ represents the weight of the $p^{th}$ feature, and $b$ is a control parameter of these weights. The membership function $u_{jl}$ must verify the constraints $\{u_{jl} \mid u_{jl} \in$



(a) (b)

Figure 3.2 – Example of data point observations distributed in bi-dimensional subspace: (a) the full dataset $X$ ($N = 4000$) versus (b) a small simple random sample points $X_{SRS}$ ($n = 40$).



(a) (b)

Figure 3.3 – (a) The within-cluster dispersion $\varphi_p$ and (b) the final feature weights $w_p$ obtained as output of W-$k$-means clustering on the dataset $X$ shown in figure 3.2 (a). The $\varphi_p$ and the $w_p$ are shown inversely proportional for the two features of the dataset $X$.

$[0, 1]$ and $\sum_{l=1}^{L} u_{jl} = 1\}$ and the feature weight $w_p$ must verify the constraints $\{w_p | w_p \in [0, 1]$ and $\sum_{p=1}^{P} w_p = 1\}$. Then, the first order condition allows writing the membership degrees, the centers and the feature weights as presented in table 3.2. Figure 3.2 (a) represents an example of dataset $X$ to be regrouped into two clusters. The dispersion of each cluster is large for the feature 1 and small for the feature 2. W-$k$-means clustering of the dataset $X$ produces as output the within-cluster dispersion $\varphi_p$ and the corresponding feature weights $w_p$ shown respectively in figures 3.3 (a) and (b). The figures show that the two parameters $\varphi_p$ and $w_p$ are inversely proportional. Thus, it is shown that $w_2 \gg w_1$ (i.e. $\varphi_2 \ll \varphi_1$) which implies that W-$k$-means promotes feature 2 than feature 1. However, figure 3.2 (a) shows visually that feature 1 is clearly more discriminative than feature 2. Therefore, minimizing only the within-cluster dispersion criteria does not necessary correspond to discriminative features. The W-$k$-means does not incorporate the global-data aspect into the clustering process and seems only insufficient to identify the relevant features from the dataset. Our approach consists to overcome the limitations of W-$k$-means by introducing a second criteria into the objective function called global-data dispersion.

# 3 The proposed approach

## 3.1 SS-$k$-means clustering

In data clustering, there is no reason to consider that features with equal importance which it will lead a more significant clustering results. Traditional $k$-means clustering techniques deal with all features equally in deciding the cluster memberships of data points. However, this is not desirable in THz imaging where pixels often contains a huge number of diverse features. The structure of the clusters in a given THz image is often restricted to a subset of features rather than the whole set of features. This leads us to ask the following questions: Is there a useful way to reduce the features space related to the structure of the clusters? Is it possible to identify the relevant features for a given pixel? In this section, we start by presenting our feature weighting formulation into the clustering process to overcome the limitations of traditional $k$-means techniques. Our approach is called pixel sampling and feature selection in $k$-means clustering (SS-$k$-means) and consists to combine the feature weighting and the simple random sampling into $k$-means clustering to provide the best tradeoff between the clustering accuracy and the low computational cost. The main idea of SS-$k$-means is to find a feature space in which the clusters are better separated. In other words, each cluster must possess a minimal dispersion, while the global data must be characterized by maximal dispersion. More formally, the dispersion within the clusters is defined by:

$$\varphi_p = \sum_{l=1}^{L} \sum_{j=1}^{n} u_{jl}^{a} d(x_{jp}, m_{lp}), \tag{3.2}$$

where $n$ represents the size of the SRS sample; $m_{lp}$ is the center of the $l^{th}$ cluster for the $p^{th}$ feature; $u_{jl}$ is the membership degree of the $j^{th}$ point in the $l^{th}$ cluster; $d(x_{jp}, m_{lp}) = (x_{jp} - m_{lp})^2$ is the distance metric that measures the similarity between a data point and a cluster center for the $p^{th}$ feature, and $a > 1$ is the fuzzification degree. This criterion must be minimized to promote features having the best compactness inside clusters. The second criterion is called *global-data dispersion* criterion, represented by $\psi_p$ and defined as follows:

$$\psi_p = \sum_{j=1}^{n} d(x_{jp}, m_p) \tag{3.3}$$

where $m_p$ is the arithmetic mean of the SRS sample for the $p^{th}$ feature. This criterion must be maximized to identify discriminative features which encourage the centers to be separated as much as possible.

Let us consider $w = (w_1, ..., w_P)$ be the weights for the $P$ features and $b$ a control parameter of these weights ($b$ must be different to 1). A compromise between minimizing $\varphi_p$ and maximizing $\psi_p$ leads to propose minimizing the following objective function:

$$J(\Phi, \Psi, w) = \sum_{p=1}^{P} w_p^b \frac{\varphi_p}{\psi_p^c}, \tag{3.4}$$

where $\Phi = (\varphi_1, \cdots, \varphi_P)$ and $\Psi = (\psi_1, \cdots, \psi_P)$ are two vectors of $P$ variables. The parameter $c$ is a real which consists to control the effect of $\psi_p$ regarding to $\varphi_p$. By an optimization way, the objective function $J(\Phi, \Psi, w)$ can be minimized with respect to $u_{jl}$, $m_{lp}$ and $w_p$. The values of the membership function $u_{jl}$ must verify the constraints $\{u_{jl} \mid u_{jl} \in [0, 1] \text{ and} \sum_{l=1}^{L} u_{jl} = 1\}$. While the values of the feature weight $w_p$ must verify the constraints $\{w_p | w_p \in [0, 1] \text{ and} \sum_{p=1}^{P} w_p = 1\}$. Then, the first order condition allows writing the membership degrees, the centers and the feature weights as follows:

$$u_{jl} = \left( \sum_{h=1}^{L} \left( \frac{\sum_{p=1}^{P} \frac{w_p^b}{\psi_p^c} d(x_{jp}, m_{lp})}{\sum_{p=1}^{P} \frac{w_p^b}{\psi_p^c} d(x_{jp}, m_{hp})} \right)^{\frac{1}{a-1}} \right)^{-1}, \tag{3.5}$$

$$m_{lp} = \frac{\sum_{j=1}^{n} u_{jl}^{a} x_{jp}}{\sum_{j=1}^{n} u_{jl}^{a}} \tag{3.6}$$

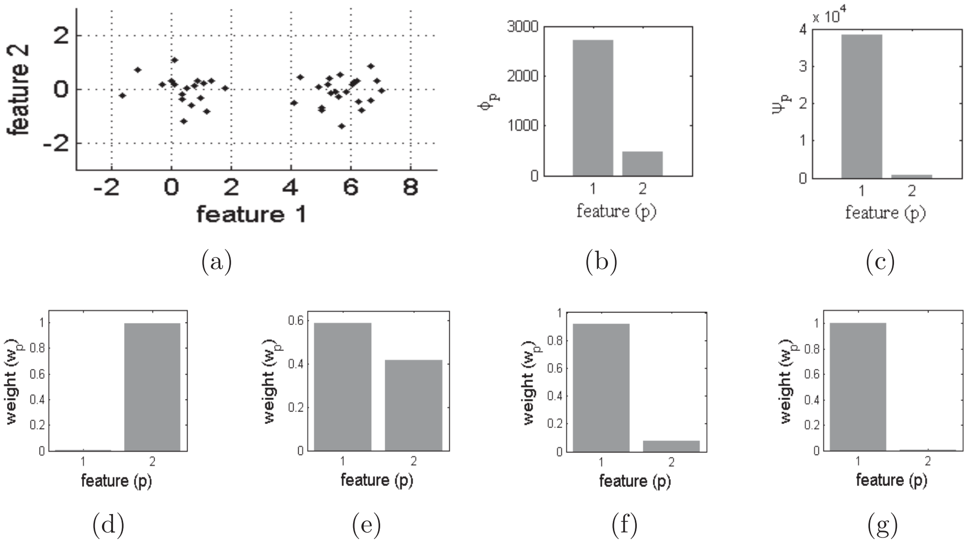
and

$$w_p = \frac{\left(\frac{\psi_p^c}{\varphi_p}\right)^{1/(b-1)}}{\sum_{t=1}^{P} \left(\frac{\psi_t^c}{\varphi_t}\right)^{1/(b-1)}}. \tag{3.7}$$

A representative sample set from the observed population is then randomly selected in order to learn the cluster centers and the feature weights. We assume that the number $L$ of clusters is known. The L-step (learning step) of SS-$k$-means consists to classify the $X_{SRS}$ into $L$ clusters of pixels; each cluster is represented by one center $m_l$ and each pixel is characterized by $P$ features and their weights $w$. The learning process is then done by iterating between three steps, updating the centers of the clusters, the membership of objects and the weights of features, until convergence, i.e. when the value of the objective function is minimized. Let us consider a parameter $Q$ inferior or equal to $P$. The $Q$ highest scores $w^*$ are identified, the corresponding features are selected, and the dimensionality of the whole set of pixels are then reduced. The L-step algorithm of SS-$k$-means can be summarized as follows:

```
L-step algorithm
1. Data: $X_{SRS}$ = SRS($X$,$n$) algorithm
   Initialize $m_l$ by random points from $X_{SRS}$
2. Do
      Update centers $m_{lq}$ using Eq. (3.6)
      Update membership degrees $u_{jl}$ using Eq. (3.5)
      Update feature weights $w_q$ using Eq. (3.7)
   Until $|J^t - J^{t-1}|$ <threshold
3. Identify the $Q$ highest weights $w_p$ and select the corresponding features. Let
   us denote $w^*$, the vector of the selected feature weights.
```

The C-step (clustering step) of SS-$k$-means consists therefore to assign each data point from the whole observed population, described in the space of the selected features, to the nearest cluster, represented by its estimated center $m_l^*$. For that reason, we propose to estimate the membership degree of data points by minimizing the objective function $J(\Phi^*, \Psi^*, w^*)$ in equation 3.4 where $w^*$ is a parameter estimated



(a) (b) (c)

(d) (e) (f) (g)

Figure 3.4 – (a) SRS sample ($X_{SRS}$) of $n = 40$ points. The SS-$k$-means clustering of $X_{SRS}$ gives the within-cluster dispersion $\varphi_p$ and the global-data dispersion $\psi_p$ shown respectively in (b) and (c). (d), (e), (f) and (g) represent the final feature weights for $c$ equal to 0, 0.5, 1 and 2.

in the L-step. The functions $\Phi^*$ and $\Psi^*$ are described in the space of the selected features associated to the whole observed population. The membership degrees of the observed data to the clusters are given by equation 3.5 using $Q$, $w^*$, $\psi_q^*$ and $m^*$ instead of $P$, $w$, $\psi_q$ and $m$. The resulted clusters are defined by the obtained membership degrees of data points.

Figure 3.4 (a) shows an example of $X_{SRS}$ sample distributed in two clusters and randomly drown from the population $X$. The $X_{SRS}$ sample represents only 1% of the observed population $X$ in figure 3.2. We propose to cluster the population $X$ by using the SS-$k$-means clustering. Figures 3.4 (b) and (c) show respectively the resulted $\varphi_p$ and $\psi_p$ associated to $X_{SRS}$. These figures show that $\varphi_1 > \varphi_2$ and $\psi_1 \gg \psi_2$. SS-$k$-means consists to promote features having a compromise between minimal values of within-cluster dispersion which corresponds to $\varphi_2$ and maximal values of global-data dispersion which corresponds to $\psi_1$. Figure 3.4 from (d) to (g) show the final feature weights ($w_1$ for feature 1 and $w_2$ for feature 2) respectively for $c$ equal to 0, 0.2, 1 and 2. For $c$ equal to 0, $w_2 > w_1$, while for 0.2, 1 and 2, $w_1 > w_2$ and $w_1$ grow when $c$ increase.

When $c > 0$, SS-$k$-means promotes then feature 1 than feature 2 which well explains the visual repartition of data. The example shows the interest of assigning weights to features by using a compromise between within-cluster and global-data dispersions associated only to a small number of data points. The computational complexity of SS-$k$-means algorithm is $O(PnL^2 + P^2)$ for one iteration, where $n$ is the sample size, $P$ the number of features and $L$ the number of clusters. This complexity is linear for parameter $n$ and quadratic for parameters $L$ and $P$. Let us recall that $k$-means [24, 77], W-$k$-means [64, 86] and SRS-$k$-means [15] are particular cases of the SS-$k$-means. The SRS-$k$-means can be obtained from equation 3.4 when $b = c = 0$, the W-$k$-means when $n = N$ and $c = 0$, and the $k$-means when $n = N$ and $b = c = 0$.

## 3.2 Sample size and feature number estimation

This section deals with the estimation of two fundamental parameters for our approach, the size $n$ of the SRS sample and the number $Q$ of the selected features. Sample size estimation is an important step in statistical sampling in which the goal is to achieve the best tradeoff between the clustering accuracy and the low computational cost. In statistics, a random sample is considered valid if it is precise. Related to reproducibility and repeatability, the SRS sample precision is the degree to which repeated sampling under unchanged conditions show the same results [104]. Let $X_{SRS} = \{x_1, x_2, \cdots, x_n\}$ is a random sample constituted by $n$ data points. We propose to randomly draw $T$ different random samples and calculate therefore their averages. Let $c_t$ represents the average or the center of the $t^{th}$ sample defined as follows:

$$c_t = \frac{1}{n} \sum_{j=1}^{n} x_j^{(t)} \tag{3.8}$$

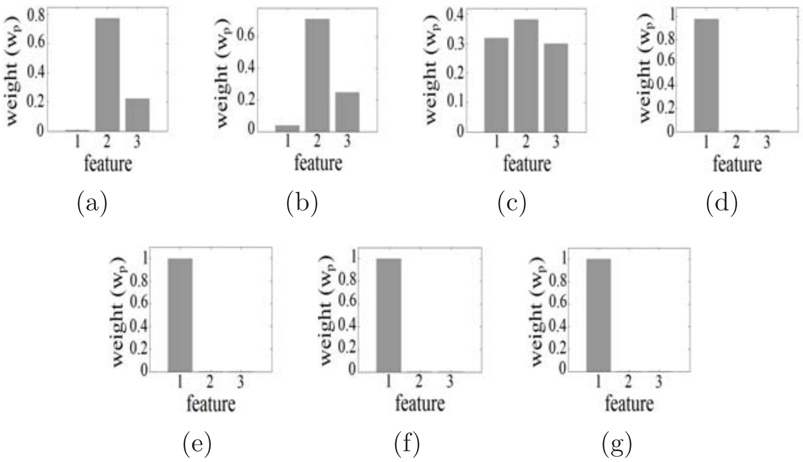
where $x_j^{(t)}$ represents a data point from the $t^{th}$ sample. The dispersion of the $T$ obtained centers $c_t$ around the mean value gives us an indicator about the precision of the random sample size. The precision can be interpreted as the closeness of agreement between independent center measurements. It is inversely proportional to a measure of the dispersion of the centers, i.e. when the precision estimator is high then the dispersion of centers is low, and inversely. The precision estimator is then defined as follows:

$$\text{precision} = -\frac{1}{T}\sum_{t=1}^{T}(c_t - \hat{c})^2 \tag{3.9}$$

where $\hat{c}$ is the mean estimator of the $T$ centers defined by $\hat{c} = \frac{1}{T}\sum_{t=1}^{T} c_t$. The precision is largely affected by random error. Increasing the sample size allows deceasing the random error and therefore increasing the sampling precision. Otherwise, we start by a minimal sample size $n = n_{min}$ units and then compute the precision estimator. We propose to increase the sample size $n$ by 1 unit and repeat these steps until the random sample size precision exceed the confidence limit CL (generally CL is equal to 95% of precisions). The corresponding algorithm can be described as follows:

```
*SampleSize*$(X, n_{min}, T)$ *algorithm*
**1.** $n \leftarrow n_{min} - 1$
**2. repeat**
        $n \leftarrow n + 1$
        **for** $t \leftarrow 1$ to $T$ **do**
            $X_{SRS}^{(t)} \leftarrow$ SRS($X$,$n$)
            Compute the center $c_t$ of $X_{SRS}^{(t)}$ using Eq. (3.8)
        **end**
        Compute the sample size precision using Eq. (3.9)
  **until** precision $>$ CL
**3. return** $n$.
```

The choice of the relevant feature number $Q$ is also an important step for our approach in which the goal is to achieve the best tradeoff between the clustering accuracy and the low computational cost. In this paper, we consider that a feature number $Q$ is valid if it is sufficient to characterize the set of pixels of a given Terahertz image. The weights of the selected features must have a sum superior to a given threshold. Let $\mathbf{w} = (w_1, w_2, \cdots, w_P)$ represents the vector of feature weights obtained by the learning step of SS-$k$-means, $w_p$ is the weight of the $p^{th}$ feature and $\sum_{p=1}^{P} w_p = 1$. Therefore, a significant number $Q$ can be computed by ranking in descending order the elements of the vector $\mathbf{w}$ and then summing the $Q$ highest weights having 85%



(a) (b) (c) (d)

(e) (f) (g)

Figure 3.5 – The final feature weights of SS-$k$-means on the synthetic dataset 1 for $c$ equal to (a) 0, (b) 0.5, (c) 1, (d) 1.5, (e) 2, (f) 2.5, and (g) 3.

of the total feature weights, defined as follows:

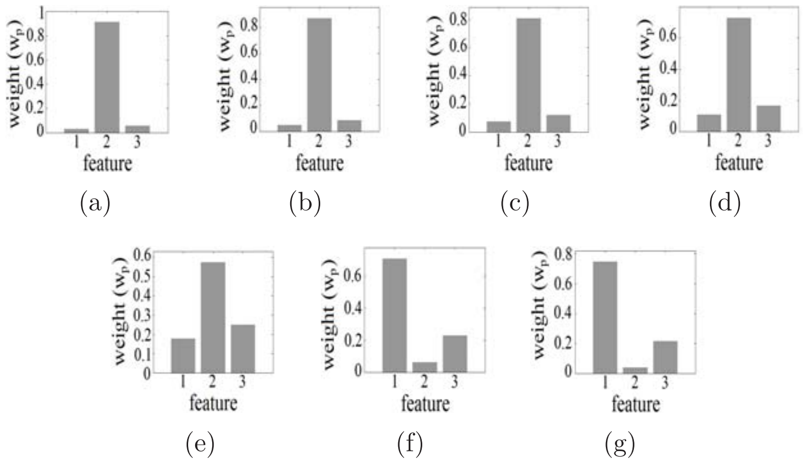
$$\sum_{q=1}^{Q} w_q \times 100 \geq 85\% \tag{3.10}$$

# 4 Experimental results

In order to study with more details the performance of our approach, synthetic datasets are used firstly to validate the clustering algorithms. Afterward, real Terahertz images are used to investigate the different approaches. Since the clustering accuracy depends on the importance of the features, we experimentally and empirically show that our approach outperforms other approaches such as traditional $k$-means, SRS-$k$-means, $k$-harmonic-means (KHM), Gaussian mixture model (GMM), and W-$k$-means.

## 4.1 Experiments on synthetic datasets

Synthetic datasets are often used to validate the clustering approaches. In this experiment, we used three synthetic datasets with different cluster number to ver-



(a) (b) (c) (d)

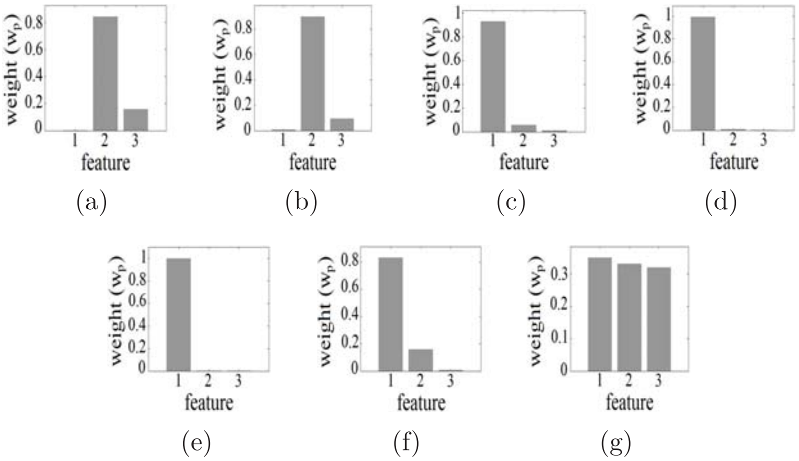
(e) (f) (g)

Figure 3.6 – The final feature weights of SS-$k$-means on the synthetic dataset 2 for $c$ equal to (a) 0, (b) 0.5, (c) 1, (d) 1.5, (e) 2, (f) 2.5, and (g) 3.

ify the performances of the clustering algorithms. The three datasets are constituted by 90000 data points and characterized by three features. Dataset 1 was generated respectively by two Gaussian distributions having means (0,0,-6) and (6,0,-6), and 3-by-3 covariance matrices containing (0.7,0.2,0.4) along the diagonal and zero off the diagonal for the two distributions. Dataset 2 was also generated respectively by two Gaussian distributions having means (0,2,0) and (0,0,2), and 3-by-3 covariance matrices containing (3,0.1,1.5) along the diagonal and zero off the diagonal for the two distributions. Dataset 3 was generated respectively by four Gaussian distributions having means (0,0,1), (-10,3,1), (10,3,1), and (20,0,-1), and 3-by-3 covariance matrices containing (0.7,0.2,0.4) along the diagonal and zero off the diagonal for the four distributions. The two parameters $a$ and $b$ were fixed to 2. The different clustering techniques were statistically compared in terms of clustering performance and silhouette cluster validity index. The clustering performance is commonly expressed as the percentage measure of the correctly clustered samples which is calculated as follows:

$$\text{Performance} = \frac{N - \text{Misclassified points}}{N} \times 100\% \tag{3.11}$$

where $N$ represents the size of the whole pixels of the Terahertz image. The clustering approaches were also compared by the silhouette cluster validity measure [96]. The



(a) (b) (c) (d)

(e) (f) (g)

Figure 3.7 – The final feature weights of SS-$k$-means on the synthetic dataset 3 for $c$ equal to (a) 0, (b) 0.5, (c) 1, (d) 1.5, (e) 2, (f) 2.5, and (g) 3.

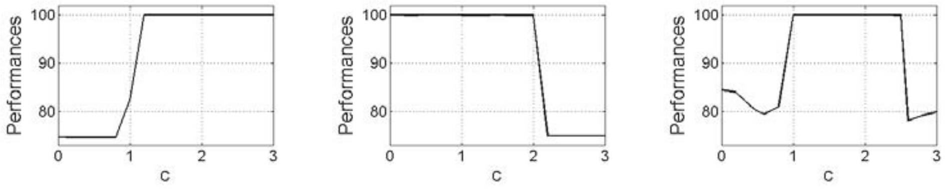
quality of clusters is measured in terms of homogeneity and separation knowing that points within one cluster are similar, while points in different clusters are dissimilar. Silhouette index is calculated as follows:

$$\text{Silhouette index} = \frac{1}{N}\sum_{i=1}^{N}\frac{\beta(x_i)-\alpha(x_i)}{\max\{\alpha(x_i),\beta(x_i)\}} \tag{3.12}$$

where $\alpha(x_i)$ is the average distance of the point $x_i$ to other points in the same cluster, $\beta(x_i)$ is the average distance of the point $x_i$ to the points in its nearest neighbor cluster. Silhouette index represents the average of the ratio in equation 3.12. A larger value indicates a better quality of the clustering result.

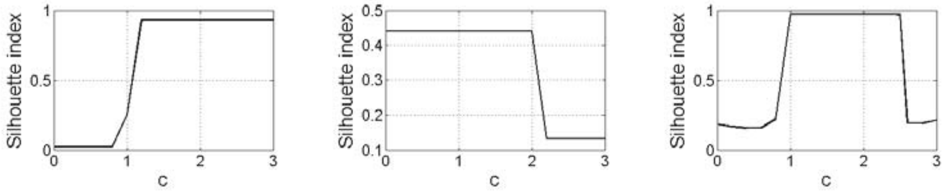
This section consists to evaluate the performances of the SS-$k$-means algorithm for seven values of the parameter $c$. The corresponding final feature weights obtained from the SS-$k$-means are shown in figures 3.5, 3.6 and 3.7 from (a) to (g) for the seven values of $c$. Figures 3.8 and 3.9 comprise three curves (for the three synthetic datasets) which represent the variation of the clustering performances and the silhouette index of the SS-$k$-means algorithm for different values of $c$. It is shown that both figures have the same variations for the three datasets.

For dataset 1, SS-$k$-means for $c = 0$ (W-$k$-means) and $c = 0.5$ promotes the second feature with a final weight near to 0.8, while the first feature has a low weight value



(a) (b) (c)

Figure 3.8 – Variation of the clustering performance provided by SS-$k$-means for different values of $c$ on (a) synthetic dataset 1, (b) synthetic dataset 2 and (c) synthetic dataset 3.



(a) (b) (c)

Figure 3.9 – Variation of the Silhouette cluster validity index provided by SS-$k$-means for different values of $c$ on (a) synthetic dataset 1, (b) synthetic dataset 2 and (c) synthetic dataset 3.

near to 0. For both cases, the clustering performances are less than 75% (figure 3.8 (a)) and the silhouette measures are near to 0 (figure 3.9 (a)). Increasing the value of $c$ to 1, the importance of the second feature decreases and its weight is sightly inferior to 0.4, while the weights of the first and the third features increase around the value 0.3 (figure 3.5 (c)). The corresponding clustering performance in figure 3.8 (a) and the silhouette measure in figure 3.9 (a) increase respectively to 83% and 0.25. Surpassing the value 1 of $c$, the corresponding feature weights are shown in figure 3.5 from (d) to (g). In that case, the first feature is promoted with a final weight near to 1, while the second and the third features have very low weight values near to 0. The clustering performances and the silhouette measures for $c > 1$ are respectively near to 100% and 1. As illustrated in figures 3.8 (a) and 3.9 (a), low values of $c$ correspond to the low performances of the SS-$k$-means clustering, while high values of $c$ correspond to the high performances.

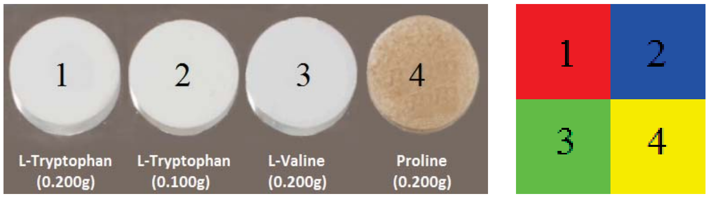


Figure 3.10 – An image of four chemical compounds acquired in visible spectrum in the left and the ground truth of the Terahertz image in the right. The false colors red, blue, green and yellow correspond respectively to the chemical compounds L-Tryptophan (0.200g), L-Tryptophan (0.100g), L-Valine (0.200g) and Proline (0.200g).

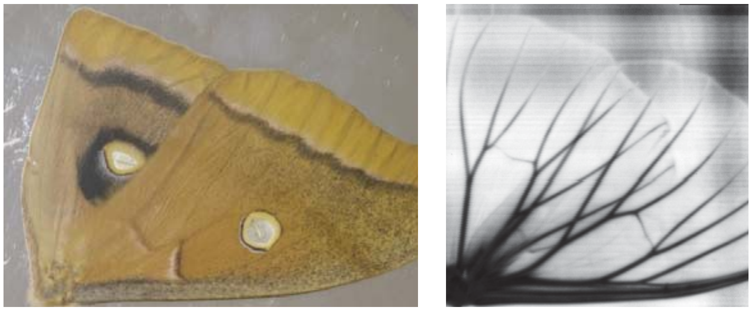


Figure 3.11 – An image of a moth acquired in visible spectrum in the left and the $570^{th}$ band of the THz image in the right.

In the case of synthetic dataset 2, the clustering performances and the silhouette measures of the SS-$k$-means are very high (respectively near to 100% and 0.44) for low values of $c$, i.e. for $c \leq 2$, (figures 3.8 (b) and 3.9 (b)). The corresponding feature weights are shown in figure 3.6 from (a) to (e). The second feature is promoted with a weight between 0.6 and 0.9, while the first and the third features have low scores which do not surpass 0.25. However, the clustering performances and the silhouette measures are low, which do not surpass the 75% and 0.13, for high values of $c$, i.e. for $c > 2$ (figures 3.8 (b) and 3.9 (b)). The resulted feature weights are shown in figures 3.6 (f) and (g). The first feature is promoted in that case with weight values around 0.75, the third feature has weight values around 0.2, while the second feature has very low weights which do not surpass 0.05.

For dataset 3, the variation of the SS-$k$-means clustering performances and silhouette measures for different values of $c$ are shown in the curve (c) of figures 3.8 and 3.9.

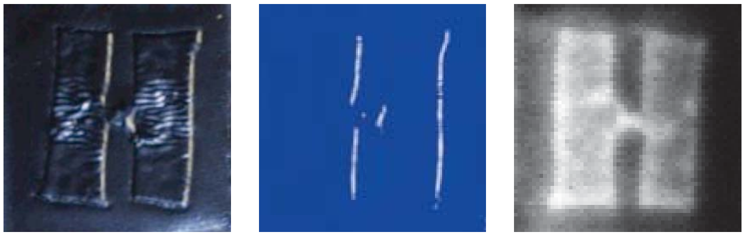


Figure 3.12 – An aluminum substrate with different thickness is acquired in the visible light. The image contains a letter "H" primed in the left and painted in the middle. The $680^{th}$ band of the paint THz image is shown in the right.

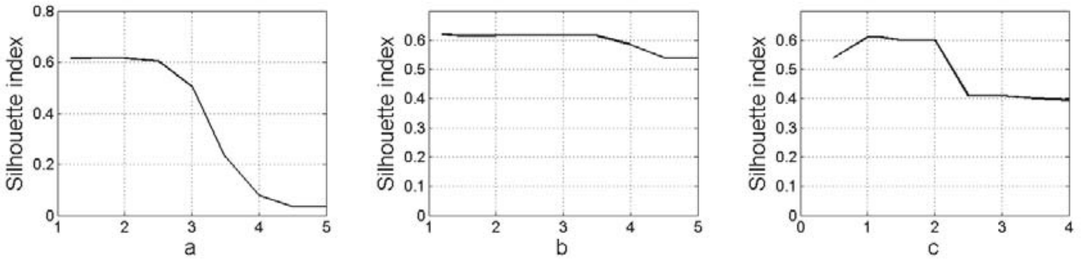


Figure 3.13 – Three curves of the silhouette index measurements of SS-$k$-means on the chemical THz image. The first curve is a function of $a$ ($b = 2$ and $c = 1.2$), the second curve is a function of $b$ ($a = 2$ and $c = 1.2$), and the third curve is a function of $c$ ($a = 2$ and $b = 2$).

On the contrary to the first two examples (dataset 1 and dataset 2), SS-$k$-means performances and silhouette measures of the synthetic dataset 3 are high (respectively near to 100% and 1) for medium values of $c$, while these measures are low (respectively inferior to 85% and 0.22) for $c < 1$ and $c > 2.5$. For $c$ between 1 and 2.5, the first feature is promoted with weight values between 0.8 and 0.9, and the second and the third features have low weight values which do not surpass 0.2 (figure 3.7 from (c) to (f)). However, for $c$ equal to 0 and 0.5, the SS-$k$-means promotes the second feature with weight values between 0.8 and 0.9, the third feature has scores inferior to 0.2, and the first feature has low scores near to 0 (figures 3.7 (a) and (b)). In the case of $c > 2.5$, the weights of the three features are approximately equals (figure 3.7 (g)). For the three datasets, we observed that suitable clustering results of SS-$k$-means can be reached with values of $c$ between 1 and 2.

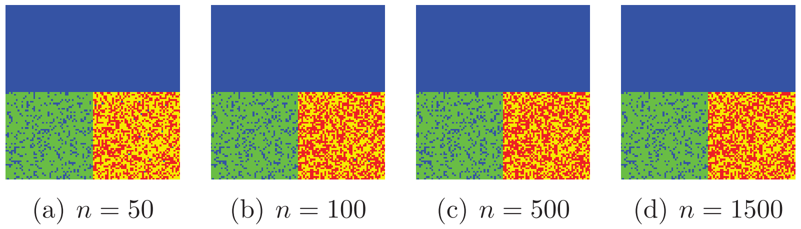
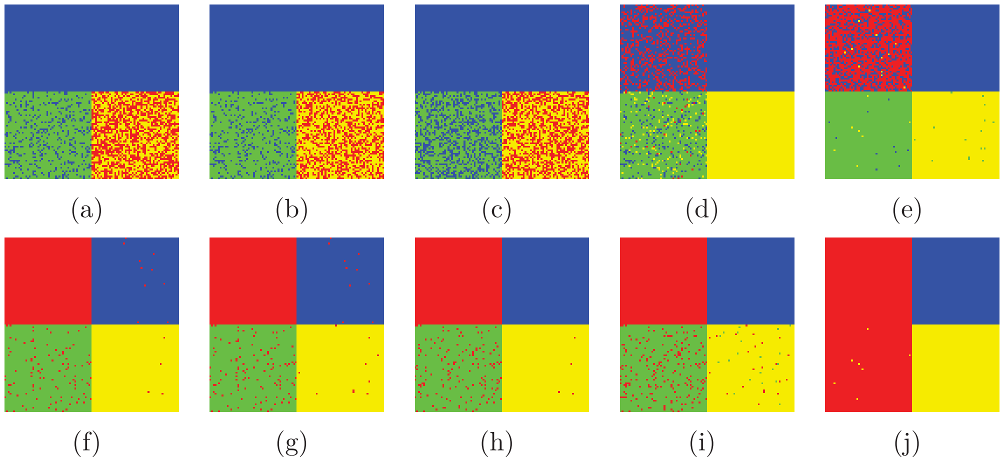

(a) $n = 50$ (b) $n = 100$ (c) $n = 500$ (d) $n = 1500$

Figure 3.14 – SRS-$k$-means segmentation of the chemical THz image for different values of the sample size $n$.

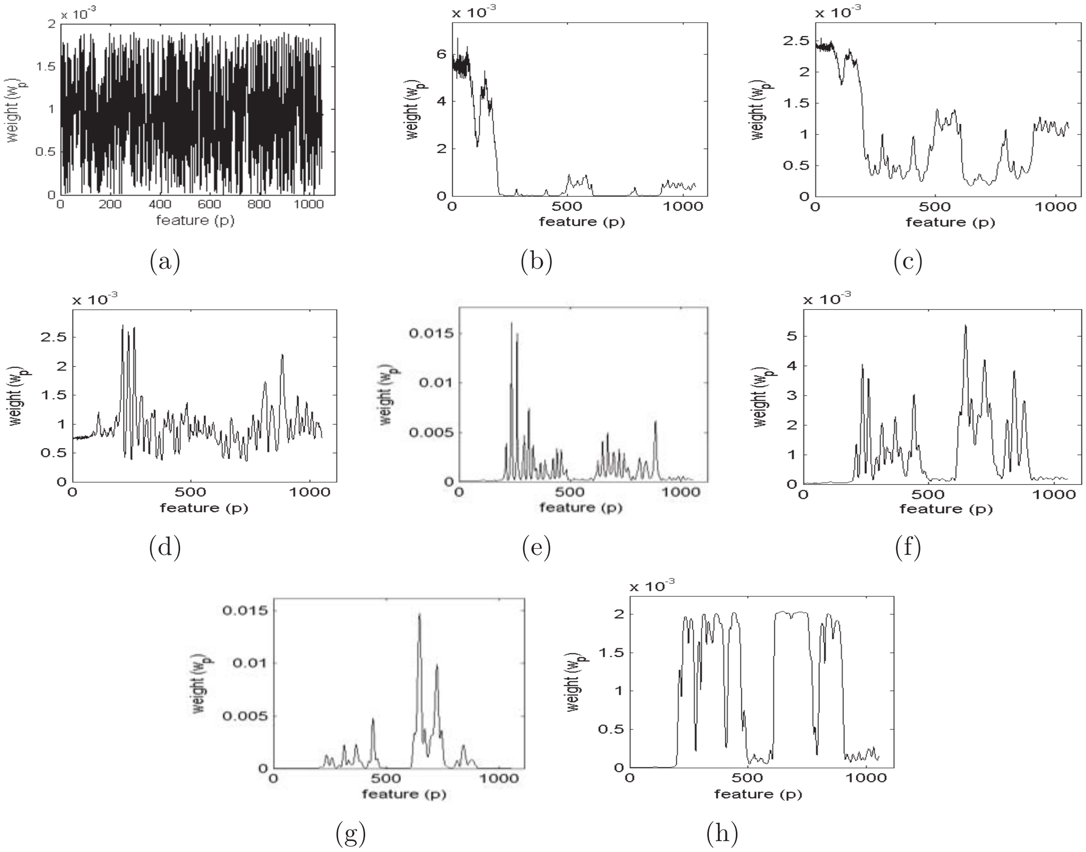

(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 3.15 – Chemical THz image segmentation for $k$-means (a), KHM (b), GMM (c), W-$k$-means (d) and SS-$k$-means for $c$ equal to 0.5 (e), 1 (f), 1.2 (g), 1.5 (h), 2 (i), and 2.5 (j).

## 4.2 Experiments on Terahertz images segmentation

In this section, SS-$k$-means, W-$k$-means, KHM, GMM, SRS-$k$-means and $k$-means are tested on chemical, moth and paint THz images. Since the THz images cannot be displayed (hundreds of bands), we present in figures 3.10, 3.11 and 3.12 the objects which were acquired in THz spectrum and used for the validation. The ground truth of the chemical THz image, the $570^{th}$ THz band of the moth image and the $680^{th}$ THz band of the paint image are shown in the right of the same figures. The chemical THz image is constituted by four compounds, L-Tryptophan (0.200g), L-Tryptophan (0.100g), L-Valine (0.200g) and Proline (0.200g), extracted and distributed into four false colored regions, the moth image is essentially constituted by two wings, and



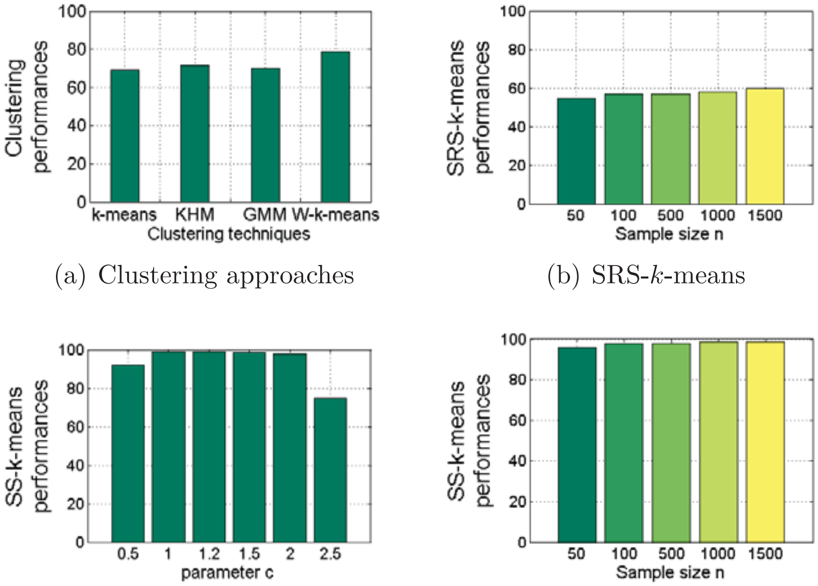
(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure 3.16 – (a) Initial random feature weights. Final feature weights of W-$k$-means (b) and SS-$k$-means on the chemical THz image for $c$ equal to 0.5 (c), 1 (d), 1.2 (e), 1.5 (f), 2 (g), and 2.5 (h).

the paint image is constituted by an aluminum substrate with different thickness which contains a letter ”H” primed and painted respectively in the left and in the middle of figure 3.12. Each pixel of chemical, moth and paint THz images are formed respectively by 1052, 894 and 1308 bands in the time domain. The feature weights and the cluster centers were randomly initialized by the same values for the different tests. The initial feature weights for chemical, moth and paint images are shown respectively in figure 3.16 (a), figure 3.22 (a) and figure 3.27 (a). The segmentation of the three images was employed respectively with 4, 5 and 3 clusters. Figure 3.13 presents the silhouette measures of SS-$k$-means on the chemical image as functions of parameters $a$, $b$ and $c$. Figure 3.13 (a) shows that a good choice of the parameter $a$ is around



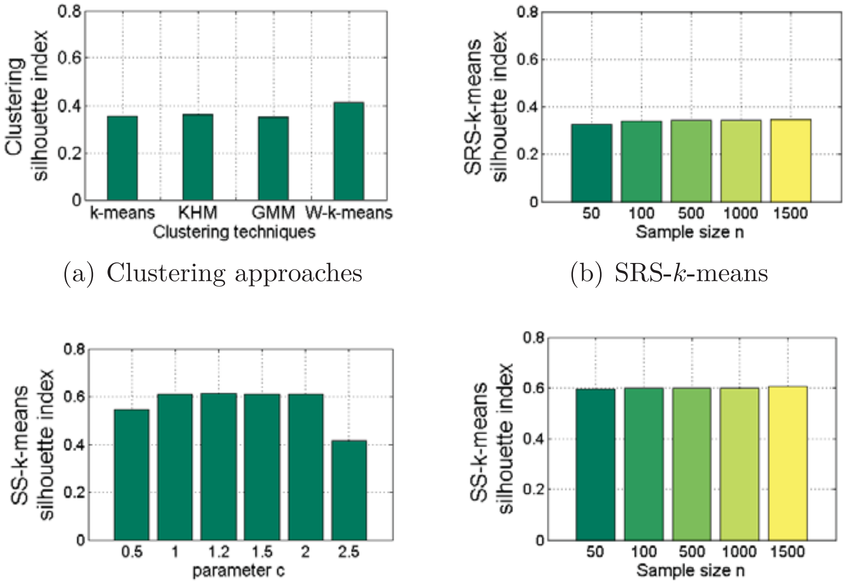
(a) Clustering approaches

(b) SRS-$k$-means

(c) SS-$k$-means (for each value of $c$)

(d) SS-$k$-means (for each value of $n$)

Figure 3.17 – Clustering performances on the chemical THz image for (a) $k$-means, KHM, GMM, W-$k$-means, (b) SRS-$k$-means, (c) SS-$k$-means for $c = 0.5$, 1, 1.2, 1.5, 2, and 2.5 ($n = N$ and $Q = P$) and (d) SS-$k$-means for $n = 50$, 100, 500, 1000 and 1500 ($c = 1.2$ and $Q = 30$).



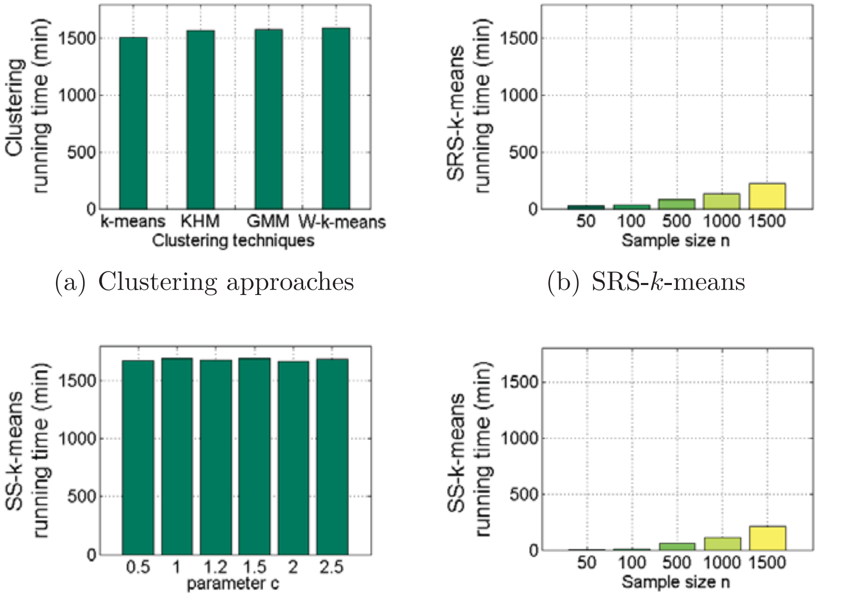
(a) Clustering approaches

(b) SRS-$k$-means

(c) SS-$k$-means (for each value of $c$)

(d) SS-$k$-means (for each value of $n$)

Figure 3.18 – Silhouette index of the chemical THz image segmentation obtained by (a) $k$-means, KHM, GMM, W-$k$-means, (b) SRS-$k$-means, (c) SS-$k$-means for $c = 0.5$, 1, 1.2, 1.5, 2, and 2.5 ($n = N$ and $Q = P$) and (d) SS-$k$-means for $n = 50$, 100, 500, 1000 and 1500 ($c = 1.2$ and $Q = 30$).



(a) Clustering approaches

(b) SRS-$k$-means

(c) SS-$k$-means (for each value of $c$)

(d) SS-$k$-means (for each value of $n$)

Figure 3.19 – Running time of the chemical THz image segmentation using (a) $k$-means, KHM, GMM, W-$k$-means, (b) SRS-$k$-means, (c) SS-$k$-means for $c = 0.5$, 1, 1.2, 1.5, 2, and 2.5 ($n = N$ and $Q = P$) and (d) SS-$k$-means for $n = 50$, 100, 500, 1000 and 1500 ($c = 1.2$ and $Q = 30$).

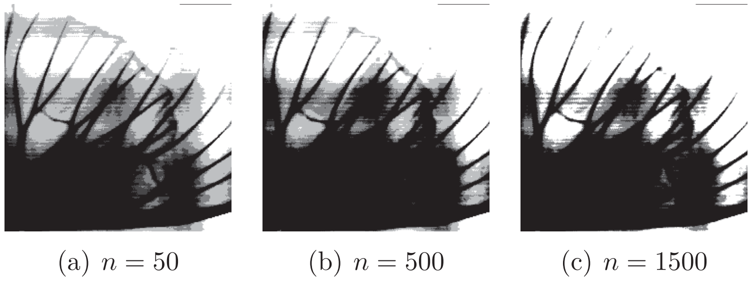


Figure 3.20 – SRS-$k$-means segmentation of the moth THz image for different values of the sample size $n$.

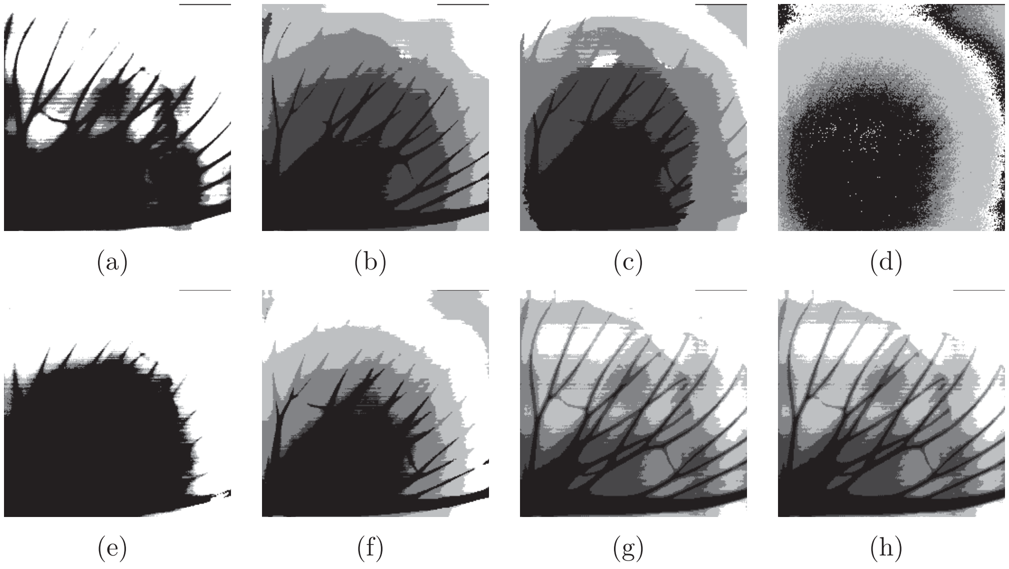
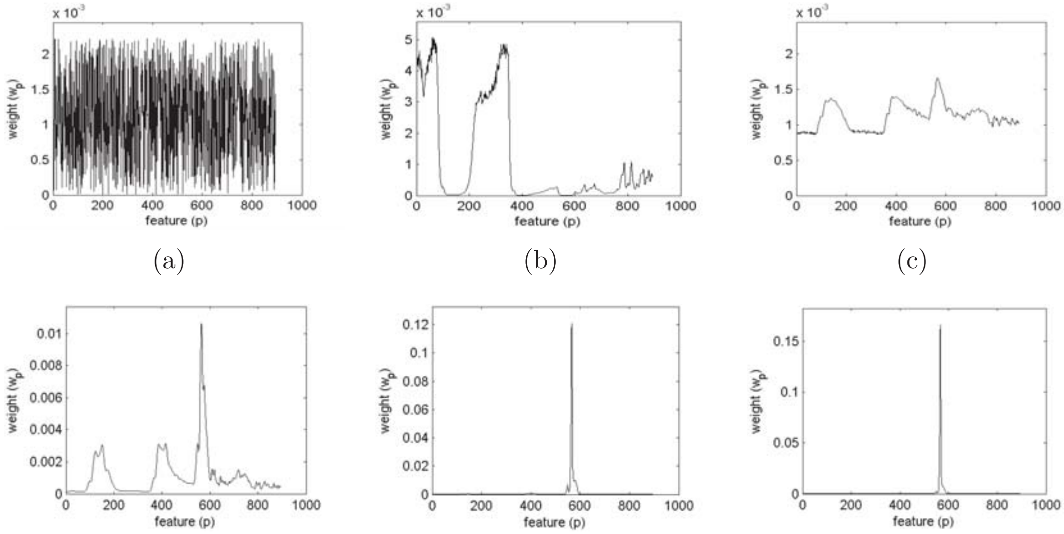


Figure 3.21 – Moth THz image segmentation for $k$-means (a), KHM (b), GMM (c), W-$k$-means (d) and SS-$k$-means for $c$ equal to 1 (e), 1.5 (f), 2 (g), and 2.5 (h).



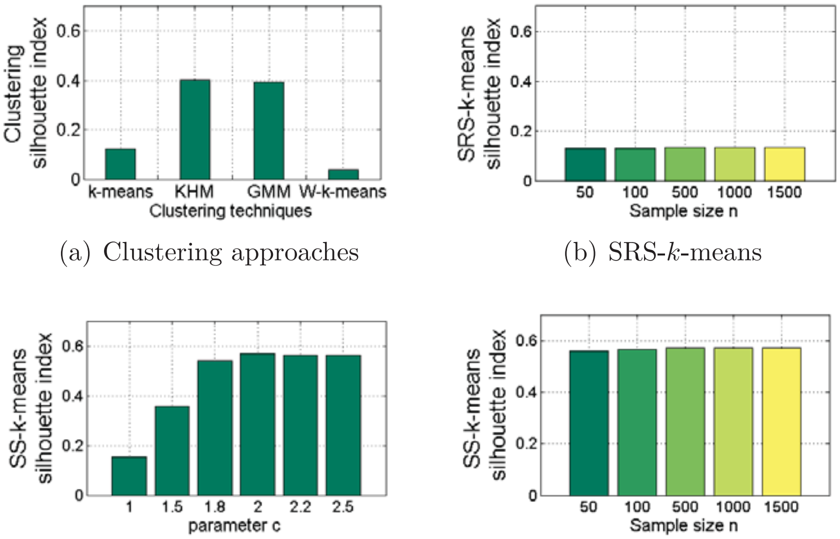
(a) (b) (c)

(d) (e) (f)

Figure 3.22 – (a) Initial random feature weights. Final feature weights of W-$k$-means (b) and SS-$k$-means on the moth THz image for $c$ equal to 1 (c), 1.5 (d), 2 (e), and 2.5 (f).

2, figure 3.13 (b) shows low effect of the values of $b$ on the clustering performances and figure 3.13 (c) shows that values between 1 and 2 of the parameter $c$ produce the highest silhouette measures. For our work, we propose to fixe the parameters $a$ and $b$ to the value 2 and study with more details the effect of the parameter $c$ on the SS-$k$-means clustering.

Figures 3.14 and 3.15 show the chemical THz image segmentation for different techniques. The SRS-$k$-means was carried out for different values of the sample size $n$ (50, 100, 500 and 1500 pixels), while the SS-$k$-means was used for $n = N$ (10000 pixels) using different values of $c$. For chemical compounds, SRS-$k$-means, $k$-means, KHM, GMM and W-$k$-means produce as output an over-segmented images (figures 3.14 (a), (b), (c) and (d) and figures 3.15 (a), (b), (c) and (d)). In the case of SRS-$k$-means (for different values of $n$), $k$-means, KHM and GMM, L-Tryptophan (0.200g) and L-Tryptophan (0.100g) clusters are fused together which clearly show their limitations using equal feature weights. However, W-$k$-means produces the final feature weights represented by the curve in figure 3.16 (b). The W-$k$-means promotes features in



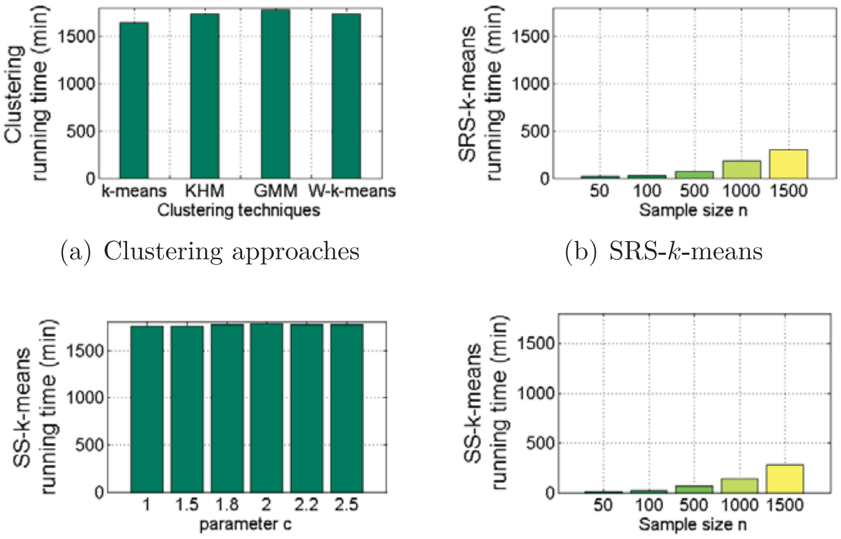
(a) Clustering approaches

(b) SRS-$k$-means

(c) SS-$k$-means (for each value of $c$)

(d) SS-$k$-means (for each value of $n$)

Figure 3.23 – Silhouette index of the moth THz image segmentation obtained by (a) $k$-means, KHM, GMM, W-$k$-means, (b) SRS-$k$-means, (c) SS-$k$-means for $c = 1$, 1.5, 1.8, 2, 2.2 and 2.5 ($n = N$ and $Q = P$) and (d) SS-$k$-means for $n = 50$, 100, 500, 1000 and 1500 ($c = 2$ and $Q = 5$).



(a) Clustering approaches

(b) SRS-$k$-means

(c) SS-$k$-means (for each value of $c$)

(d) SS-$k$-means (for each value of $n$)

Figure 3.24 – Running time of the moth THz image segmentation using (a) $k$-means, KHM, GMM, W-$k$-means, (b) SRS-$k$-means, (c) SS-$k$-means for $c = 1$, 1.5, 1.8, 2, 2.2 and 2.5 ($n = N$ and $Q = P$) and (d) SS-$k$-means for $n = 50$, 100, 500, 1000 and 1500 ($c = 2$ and $Q = 5$).

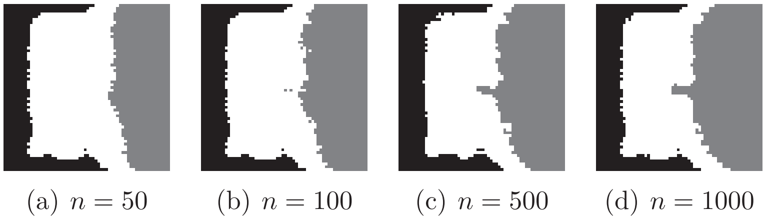


Figure 3.25 – SRS-$k$-means segmentation of the paint THz image for different values of the sample size $n$.

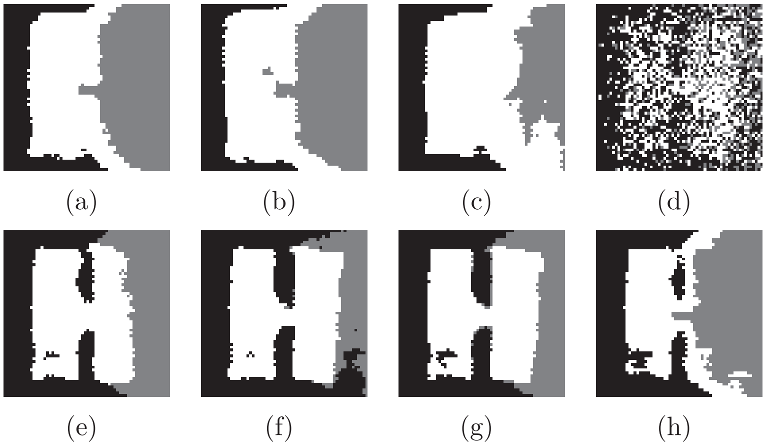


Figure 3.26 – The paint THz image segmentation for $k$-means (a), KHM (b), GMM (c), W-$k$-means (d) and SS-$k$-means for $c$ equal to 0.5 (e), 1 (f), 1.5 (g) and 2 (h).

the interval [1,200] which are not discriminative and lead to over-segmented regions. Figure 3.15 from (e) to (j) display the obtained regions of SS-$k$-means for $c$ equal to 0.5, 1, 1.2, 1.5, 2 and 2.5. For $c = 0.5$, the red region is ameliorated compared to W-$k$-means and begins to be clearly formed. The final feature weights are represented in figure 3.16 (c) and the highest scores are in the intervals [1,200], [490,680], and [910,1052]. Among them, some features are not yet discriminative to improve the clustering. The best chemical image segmentations are obtained when $c$ surpasses 0.5 in figures 3.15 (f), (g), (h) and (i), the four compounds become very well identified, except some points of L-Valine (0.200g) are misclassified. The corresponding feature weights are shown in figures 3.16 (d), (e), (f) and (g) and the pertinent bands are around 250, 425, 610 and 720. However, when $c \geq 2.5$, SS-$k$-means segmentation

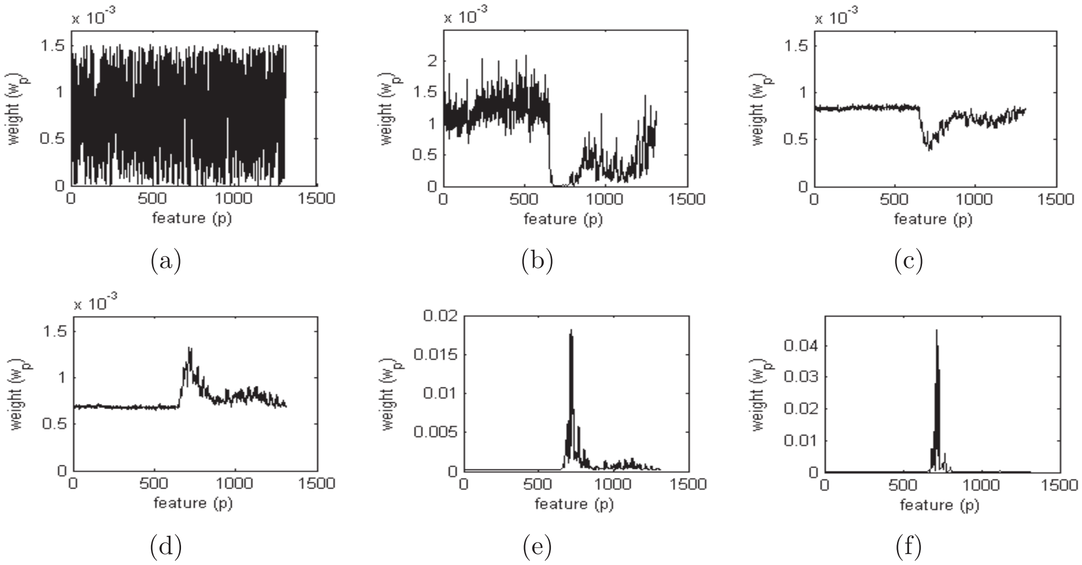
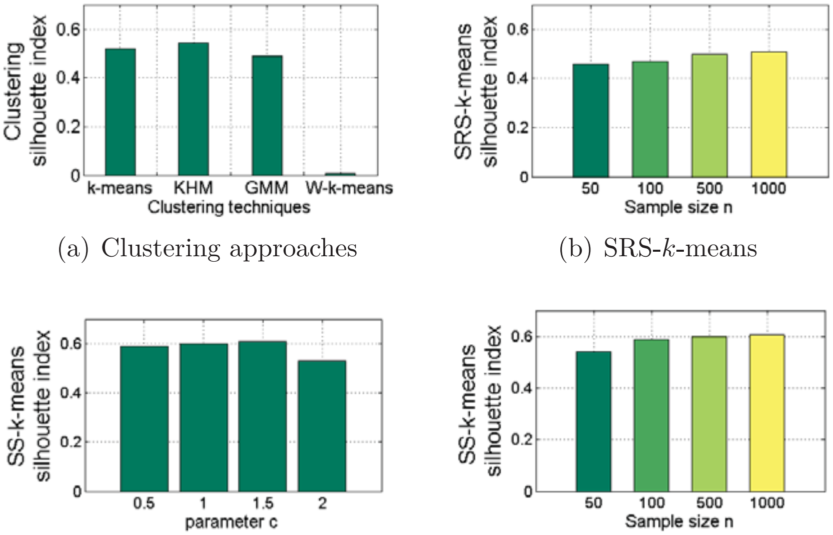


Figure 3.27 – (a) Initial random feature weights. Final feature weights of W-$k$-means (b) and SS-$k$-means on the paint THz image for $c$ equal to 0.5 (c), 1 (d), 1.5 (e) and 2 (f).

of the chemical THz image produces under-segmented regions and the red and the green regions which represent respectively the L-Tryptophan (0.200g) and the L-Valine (0.200g) are fused together. The clustering techniques ($k$-means, KHM, GMM, W-$k$-means, SRS-$k$-means and SS-$k$-means) were also statistically compared in terms of clustering performance, silhouette cluster validity index and running time. The statistics of the different approaches are shown in figures 3.17, 3.18 and 3.19. The clustering performances have not surpassed the 70% for $k$-means, KHM, GMM, and SRS-$k$-means, 80% for W-$k$-means, 93% for SS-$k$-means with $c = 0.5$, 75% for SS-$k$-means with $c = 2.5$, and near to 100% for SS-$k$-means with parameter $c$ from 1 to 2 ($n = N$). The performances of SS-$k$-means are also near to 100% for low sample size $n \ll N$ ($c = 1.2$). The silhouette validity indices are around the value 0.36 for $k$-means, KHM, GMM, and SRS-$k$-means techniques, 0.42 for W-$k$-means, 0.55 for SS-$k$-means with $c = 0.5$, 0.41 for SS-$k$-means with $c = 2.5$, 0.61 for SS-$k$-means with parameter $c$ from 1 to 2 and around 0.61 for SS-$k$-means for low values of $n$ ($c = 1.2$). The running time is lower for SS-$k$-means and SRS-$k$-means with small sample size $n$ ($n \ll N$) and higher for $k$-means, KHM, GMM, W-$k$-means, and SS-



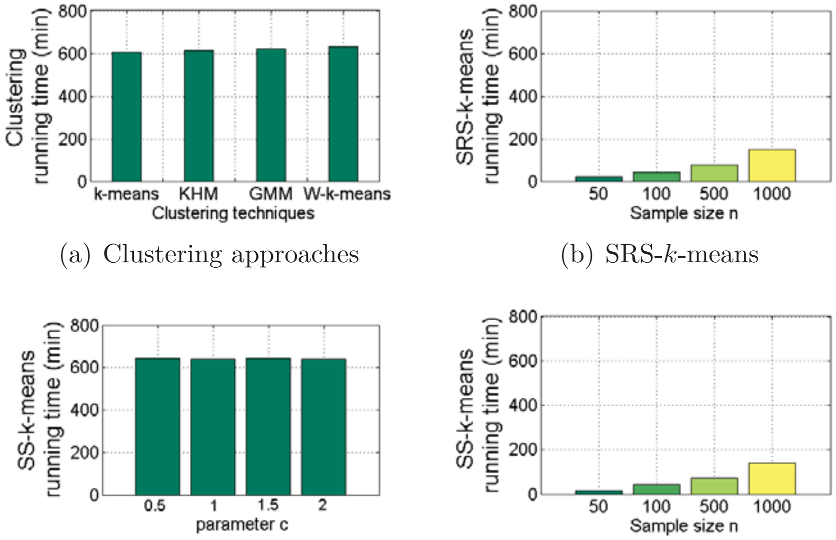
(a) Clustering approaches

(b) SRS-$k$-means

(c) SS-$k$-means (for each value of $c$)

(d) SS-$k$-means (for each value of $n$)

Figure 3.28 – Silhouette index of the paint THz image segmentation obtained by (a) $k$-means, KHM, GMM, W-$k$-means, (b) SRS-$k$-means, (c) SS-$k$-means for $c = 0.5$, 1, 1.5 and 2 ($n = N$ and $Q = P$) and (d) SS-$k$-means for $n = 50$, 100, 500 and 1000 ($c = 1.5$ and $Q = 100$).



(a) Clustering approaches (b) SRS-$k$-means

(c) SS-$k$-means (for each value of $c$) (d) SS-$k$-means (for each value of $n$)

Figure 3.29 – Running time of the paint THz image segmentation using (a) $k$-means, KHM, GMM, W-$k$-means, (b) SRS-$k$-means, (c) SS-$k$-means for $c = 0.5$, 1, 1.5 and 2 ($n = N$ and $Q = P$) and (d) SS-$k$-means for $n = 50$, 100, 500 and 1000 ($c = 1.5$ and $Q = 100$).

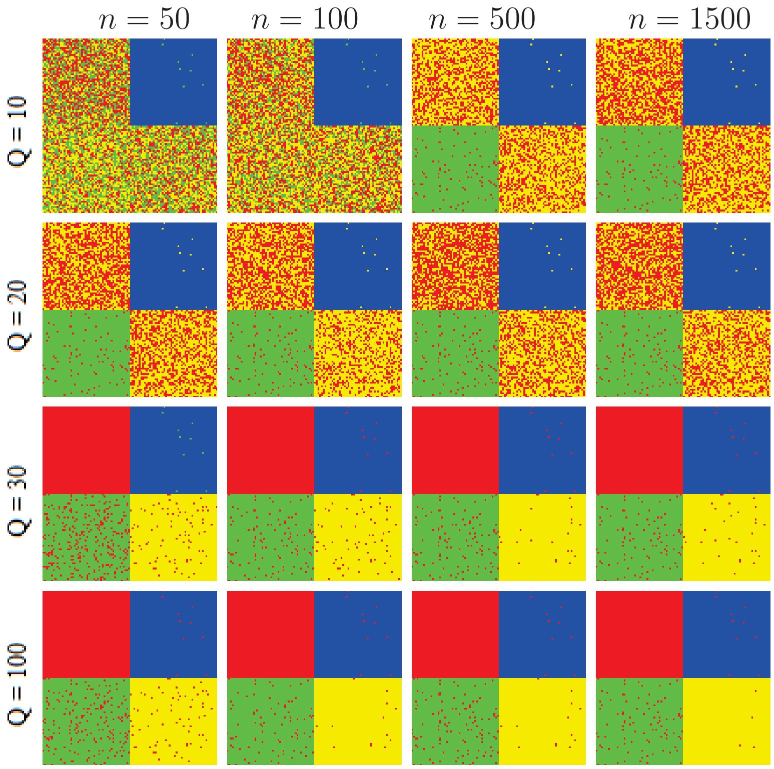
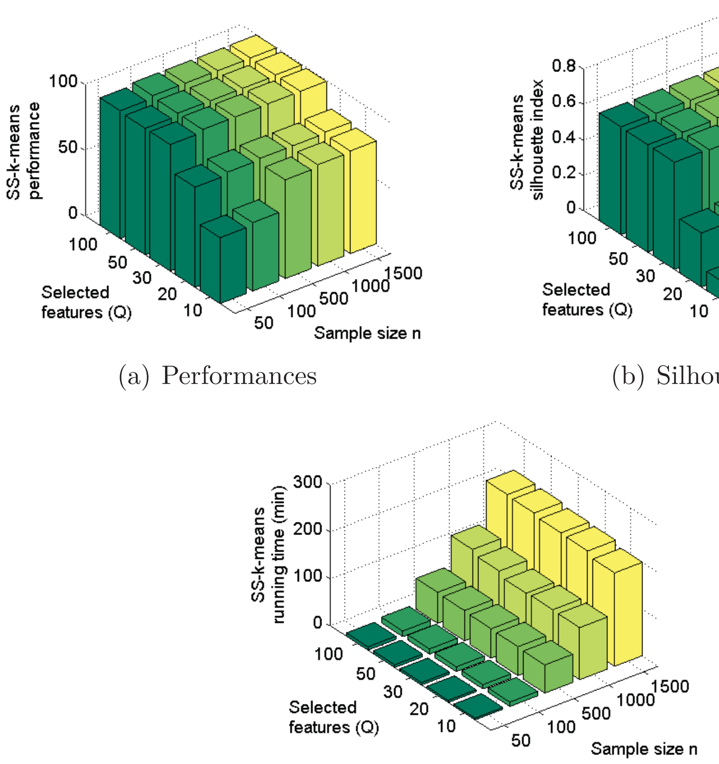


Figure 3.30 – The chemical THz image segmentation of SS-$k$-means for different values of $n$ and $Q$.

$k$-means ($n = N$) because the use of the whole pixels and features into the clustering process occupies high time and memory consuming. The obtained statistics confirm the results previously illustrated and show the high performances of our approach for low sample size.

Figures 3.20 and 3.21 show the segmentation outputs of the six clustering algorithms on the moth THz image. The $k$-means, SRS-$k$-means, KHM, GMM and W-$k$-means produce poor segmented regions in figure 3.20 and figure 3.21 from (a) to (d). The obtained regions clearly illustrate the limitations of these techniques to provide good structure of wings. Figure 3.22 (b) shows that the feature weights estimated by using W-$k$-means in the intervals [1,100] and [220,380] are not relevant, which leads to the under-segmentation. Figure 3.21 from (e) to (h) display the obtained regions of SS-$k$-means for $c$ equal to 1, 1.5, 2, and 2.5. The structure of the moth wings is again unfavorably segmented for $c$ equal to 1 and 1.5. The corresponding highest feature



(a) Performances

(b) Silhouette index

(c) Running time

Figure 3.31 – Clustering performances, Silhouette index, and running time of SS-$k$-means on the chemical THz image for different values of $n$ and $Q$.

weights in figures 3.22 (c) and (d) are around 150, 400, and 680. Among them, some features are not relevant which explain the decrease of the clustering performances. The best regions are obtained when $c$ surpass 1.5 in figures 3.21 (g) and (h). The structure of the moth wings are preserved. The corresponding feature weights are shown in figures 3.22 (e) and (f) and the pertinent bands are around 580. Visual results of the moth Terahertz image is supported by the statistics shown in figures 3.23 and 3.24. As the ground truth of the moth THz image is not very precise in our work, we study and show the performances of the clustering techniques only in terms of silhouette cluster validity index and running time. The silhouette validity indices are around the value 0.12 for the $k$-means and the SRS-$k$-means, 0.03 for the W-$k$-means, around 0.4 for the KHM and the GMM, 0.17 for SS-$k$-means with $c = 1$,

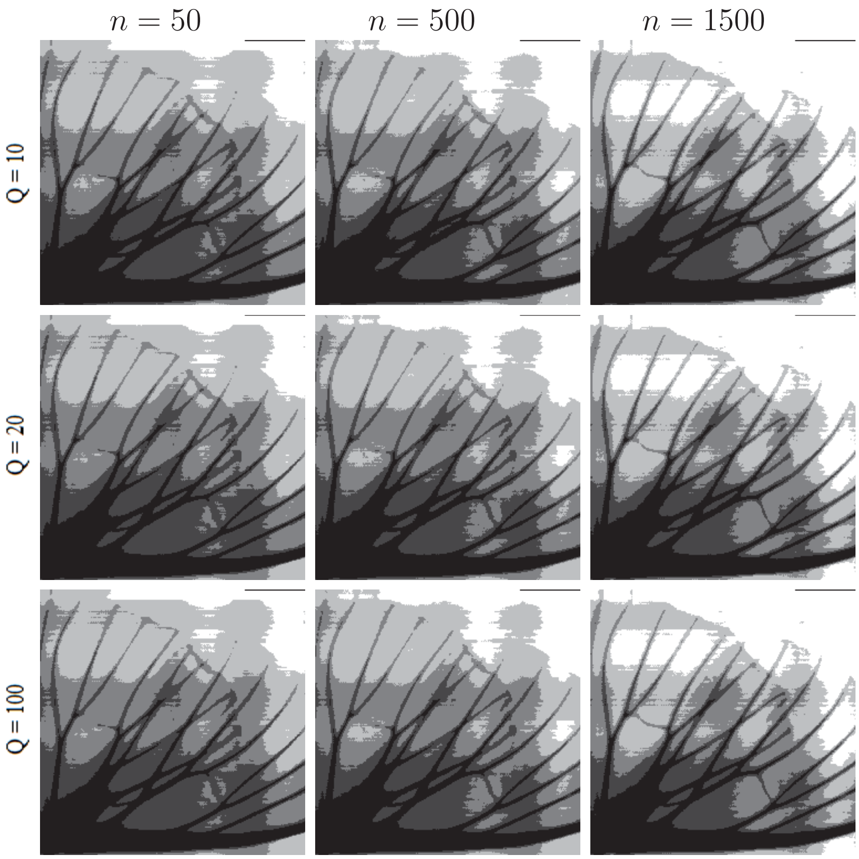
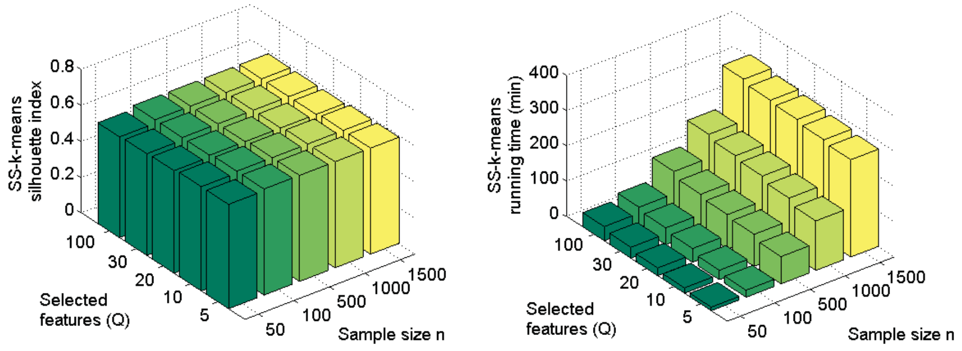


Figure 3.32 – The moth THz image segmentation of SS-$k$-means for different values of $n$ and $Q$.



(a) Silhouette index

(b) Running time

Figure 3.33 – Silhouette index and running time of SS-$k$-means on the moth THz image for different values of $n$ and $Q$.

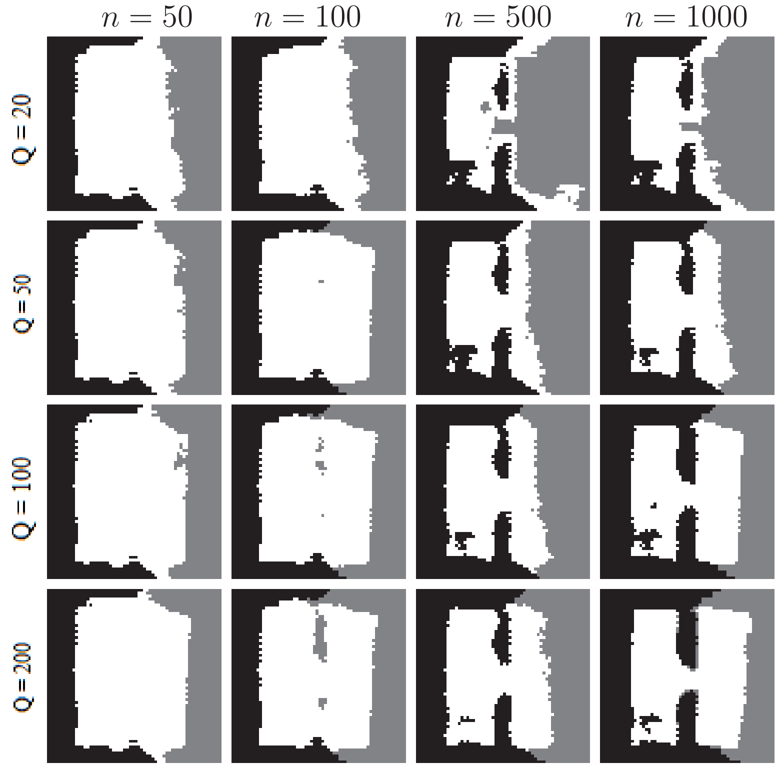
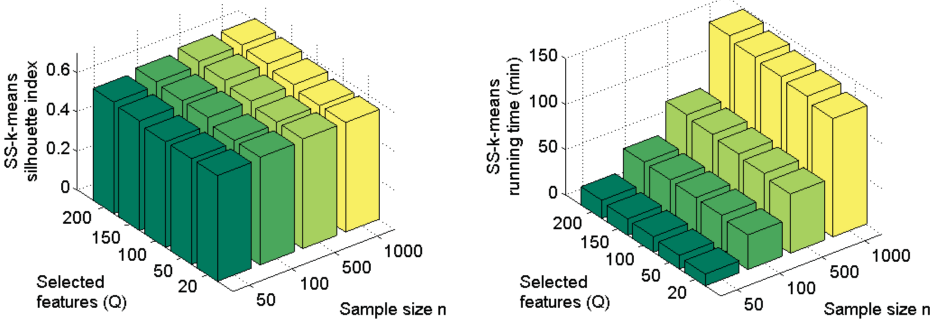


Figure 3.34 – The paint THz image segmentation of SS-$k$-means for different values of $n$ and $Q$.

0.37 for SS-$k$-means with $c = 1.5$, 0.56 for SS-$k$-means with parameter $c > 1.5$, and around 0.56 for SS-$k$-means for low sample size $n \ll N$ $(c = 2)$. The running time is again lower for SS-$k$-means and SRS-$k$-means with small sample size $n$ and higher for $k$-means, KHM, GMM, W-$k$-means, and SS-$k$-means for each value of $c$ and for $n = N$. The obtained statistics confirm the results previously illustrated especially for low sample size.

Figures 3.25 and 3.26 show the segmentation outputs of the different algorithms on the paint THz image. The $k$-means, SRS-$k$-means, KHM, GMM and W-$k$-means produce poor segmented regions in figure 3.25 and figure 3.26 from (a) to (d). The obtained regions clearly illustrate the limitations of these techniques to discover the true shape of the letter H. Figure 3.27 (b) shows that the feature weights estimated by using W-$k$-means are in the interval [1,630]. The features having the high scores are not relevant and leads to the poor segmentation results. Figure 3.26 from (e)



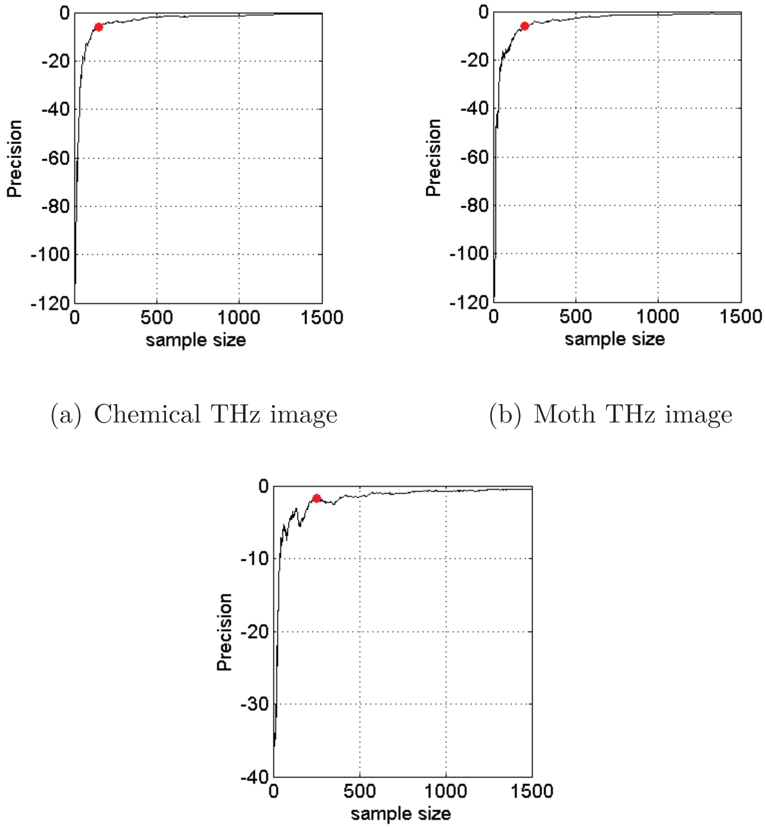
(a) Silhouette index

(b) Running time

Figure 3.35 – Silhouette index and running time of SS-$k$-means on the paint THz image for different values of $n$ and $Q$.

to (h) display the obtained regions of SS-$k$-means for $c$ equal to 0.5, 1, 1.5 and 2. The structure of the image is favorably segmented for $c$ equal to 0.5, 1 and 1.5. The pertinent bands (figures 3.27 (c), (d) and (e)) are around 580. Visual results of the paint Terahertz image is supported by the statistics shown in figures 3.28 and 3.29. As the ground truth of the paint THz image is not available in our work, we study and show the performances of the clustering techniques only in terms of silhouette cluster validity index and running time. The silhouette validity indices are around the value 0.52 for the $k$-means, the SRS-$k$-means, the KHM and the GMM, 0.02 for W-$k$-means, around 0.6 for SS-$k$-means with $c$ equal to 0.5, 1 and 1.5, near to 0.53 for SS-$k$-means with parameter $c = 2$, and around 0.6 for SS-$k$-means for low sample size $n \ll N$ ($c = 1.5$). The running time is again lower for SS-$k$-means and SRS-$k$-means with small sample size $n$ and higher for $k$-means, KHM, GMM, W-$k$-means, and SS-$k$-means for each value of $c$ and $n = N$. The obtained statistics confirm the results previously illustrated and show that our approach is accurately more rapid when it is used with a small sample size.

In figure 3.30, the output regions of SS-$k$-means segmentation (for $c = 1.2$) of the chemical THz image is shown for different values of $n$ and $Q$. Note that when $Q$ surpassing 20 and for small sample size $n$, the results are very interesting. Figures 3.31 (a), (b) and (c) present the statistics of SS-$k$-means in terms of clustering performances, silhouette measures and time running. The obtained statistics confirm the



(a) Chemical THz image

(b) Moth THz image

(c) Paint THz image

Figure 3.36 – Variation of the precision of different sample sizes for chemical, moth and paint Terahertz images.

efficiency of SS-$k$-means segmentation for low values of $n$ and $Q$ and show its rapidity compared with the other approaches. As detailed in section 3.2, the parameters $n$ and $Q$ can be estimated respectively for high sample size precision and highest weights of relevant features. Figures 3.36 (a) shows a curve representing the precision variations for different values of $n$. The optimal sample size $n$ are found equal to 148 pixels for $CL = -5.60$ (represented by the red disk in figure 3.36 (a)). While, the significant number $Q$ of relevant features are found equal to 53. Therefore, a small random sample of pixels, around 1%, and a minimal number of features, around 5%, are sufficient to produce favorable segmentation using SS-$k$-means.

In figure 3.32, the SS-$k$-means segmentation (for $c = 2$) of the moth THz image is shown for different values of $n$ and $Q$. The regions output outperform those obtained by the other clustering approaches. Figures 3.33 (a) and (b) show the statistics of the SS-$k$-means in terms of silhouette index and time running. The obtained statistics confirm the efficiency of SS-$k$-means and show its rapidity compared to $k$-means, KHM, GMM and W-$k$-means techniques. Figure 3.36 (b) shows a curve representing the variations of precisions for different values of $n$. The optimal sample size $n$ are found equal to 194 pixels for $CL = -5.88$ (represented by the red disk in figure 3.36 (b)). While the number $Q$ of relevant features are found equal to 21 features.

In figure 3.34, the SS-$k$-means segmentation (for $c = 1.5$) of the paint THz image is shown for different values of $n$ and $Q$. The regions output outperform those obtained by the other clustering approaches. Figures 3.35 (a) and (b) show the statistics of the SS-$k$-means in terms of silhouette index and time running. The obtained statistics confirm the efficiency of SS-$k$-means and show its rapidity compared to $k$-means, KHM, GMM and W-$k$-means techniques. Figure 3.36 (c) shows a curve representing the variations of precisions for different values of $n$. The optimal sample size $n$ are found equal to 254 pixels for $CL = -1.9$ (represented by the red disk in figure 3.36 (c)). While the number $Q$ of relevant features are found equal to 78 features. Note that only around 2% of pixels and 5% of features are sufficient to segment accurately the three THz images.

## 5 Conclusion

In this paper, we have proposed a novel clustering approach, called SS-$k$-means, to segment THz images. Feature weighting is used in order to reduce the number of features required for carrying out the segmentation. In addition to the computational time, irrelevant features decreases the clustering accuracy. The SRS scheme allows to use around 1% of pixels. Automatic estimation of the sample size $n$ and the selected feature number $Q$ are also proposed in this paper. Our approach is more appropriate for achieving the best compactness inside clusters and the best discrimination of features. It is evaluated and compared favorably with some related works.

SS-$k$-means is shown so attractive to achieve the best clustering accuracy and the low computational cost by using only low sample size $n$ and low features number $Q$, especially for $c$ between 1 and 2. Note that precise choice of $c$ remains very interesting for the clustering performances. Furthermore, the sensitivity to initial starting centers and feature weights decreases the clustering accuracy. These problems haven't been addressed in this paper and require further studies.

In further work, we will deal with the estimation of the parameter $c$. For instance, the parameter $c$ can be estimated by writing equation 3.4 in the form of probability and adding a prior on $c$. Similar steps to the work of Allili and Ziou [8] about the variational calculus, one can obtain a value about the parameter $c$.

# Acknowledgments

The authors would like to thank Professor Thomas Tongue of Zomega THz Corporation for the THz measurements.

# Chapitre 4

# Classification $K$-Autorégressive pour la segmentation d'images THz

Dans les deux chapitres précédents, la propriété de corrélation entre les bandes de l'image Térahertz n'est pas utilisée dans le processus de classification. Dans ce chapitre, nous introduisons une nouvelle famille de techniques de classification basées sur la régression et qui sont adaptées aux séries chronologiques. Nous supposons que les valeurs associées à chaque pixel d'une image Térahertz sont échantillonnées à partir d'un modèle autorégressif. La segmentation de l'image est alors vue comme un problème de classification de séries chronologiques. La classification est formulée comme un problème d'optimisation non-linéaire. L'ordre du modèle et le nombre de classes sont automatiquement estimés en utilisant un critère de sélection de modèle.

Dans ce chapitre, nous présentons un article intitulé **$K$-Autoregres-sive clustering for Terahertz image segmentation** soumis au journal international de Elsevier **Pattern Recognition**. Le problème a été posé par le professeur Djemel Ziou. J'ai réalisé, validé et rédigé ce travail sous sa supervision. Une version compacte de ce travail a été publiée dans la conférence internationale de Springer **International Conference Image Analysis and Recognition (ICIAR2017)**, Montréal, Canada, 2017, intitulée **$K$-Autoregressive Clustering : Application on Terahertz Image Analysis** [12].

# $K$-Autoregressive clustering for Terahertz image segmentation

Mohamed Walid Ayech
Département d'informatique, Université de Sherbrooke,
Sherbrooke, Québec, Canada J1K 2R1
walid.ayech@usherbrooke.ca

Djemel Ziou
Département d'informatique, Université de Sherbrooke,
Sherbrooke, Québec, Canada J1K 2R1
djemel.ziou@usherbrooke.ca

**Keywords**: Segmentation, Terahertz imaging, time series, autoregressive model, model selection criterion.

### Abstract

Terahertz (THz) imaging provides a large amount of specific information considered as time series behind every pixel of the image. In this paper, we propose to segment THz images and introduce a new family of clustering suitable to time series. In particular, we propose a novel approach called $K$-Autoregressive ($K$-AR) model in which we assume that the time series depicting the pixels were generated by univariate AR models. $K$-AR approach consists to classify these AR models and minimize a new objective function for recovering the original $K$ autoregressive models describing each cluster of time series. The corresponding pixels are then assigned to the clusters having the best AR model fitting. The order and cluster number of $K$-AR are automatically estimated using a model selection criterion. $K$-AR is tested on various artificial datasets and Terahertz images. Experimental results show that $K$-AR is more efficient than other approaches from the literature.

# 1 Introduction

Recently, clustering of time series has attracted a lot of interest in many scientific studies [73, 45, 44, 46]. It is often essential to find solutions for real problems deriving from several application domains such as bioinformatics, environmetrics, genetics, multimedia and finance [2]. In environmetrics, for instance, time series clustering has been used to group a set of air pollutant emissions gathered at different times for inspecting the efficiency of an environmental monitoring network [48, 46]. In genetics, time series clustering has been used to group genes taking into account profiles of temporal expression from cDNA microarrays experiments [84]. In biomedicine, the study of EEG biological signals requires to discriminate between signals caused by sick or healthy people [87]. Several other applications of time series clustering are detailed in [2].

In statistics, a time series is a sequence of observations related in chronological order. It is considered as a vector represented by high number of bands or features. Several works of time series clustering are based on different kind of distance measures [73]. Among these distances, short time series (STS) distance [84], dynamic time warping (DTW) distance [73], autocorrelation based distance [48], cepstrum based distance [79] and wavelets decomposition based distance [44]. Other works use Euclidian distance based autoregressive (AR) models [48]. The clustering of the set of time series is realized in the AR parameters space. The principal advantage of this method is that the clustering process is realized in low dimensional feature space. However, the estimation of these parameters is based on the resolution of system of equation for each time series, independently. They are estimated without using information about the structure of the time series clusters. Moreover, the interpretation of these parameters and the space which constitute is not easy. For example, what is the mean of the first coefficients of the AR model? Beside, standard clustering techniques of the set of time series are based on similarity measurements (e.g. Euclidean distance) between the abstract parameters. Instead, we believe to model each cluster of time series by a single univariate AR model. The AR parameters must be estimated by using the correlation property between time series variables and also the structure of the time series clusters. We look forward an iterative algorithm which alternate between autoregressive modeling and time series clustering until reaching a high time series fitting.

In this paper, we introduce a novel clustering approach called $K$-Autoregressive ($K$-AR) model which consists to classify the set of time series into a fixed number of clusters, each one is represented by its predictive prototype. The proposed approach consists to analyse Terahertz (THz) images constituted by several pixels, each one represents an univariate time series constituted by several temporal bands (e.g. 1000 bands). We assume that each time series is stationary and generated from an univariate autoregressive (AR) model where the order is unknown. A new objective function is proposed for recovering the original autoregressive models describing each cluster of time series and then assigns the corresponding pixels to the clusters according to their predictive prototypes. The order and the number of clusters of the $K$-AR approach are automatically estimated using a modified generalized information criterion. There are five main differences between our work and the state of art of Terahertz image analysis. First, our work constitutes a new family of clustering based autoregressive model which is suitable to univariate time series. Second, a single AR model is a representative of a class which can be used in subsequent steps THz image processing such as the object recognition. Third, $K$-AR approach is more adapted to univariate time series and not excessively affected by outliers. Fourth, our approach is based on AR feature space suitable for high dimensional THz image segmentation. Fifth, our approach is completely unsupervised, the order and the number of clusters of $K$-AR model are automatically estimated by using a modified generalized information criterion.

The rest of the paper is organized as follows: in section 2, we present an insight about related works of various applications of Terahertz imaging based time series. Moreover, a background of AR modeling for univariate time series has been detailed. In section 3, we formulate the segmentation of Terahertz images by using the clustering of AR models. The selection of the order and the number of clusters of the $K$-AR model is proposed in section 4. The results are illustrated and discussed in section 5.

# 2 Background

## 2.1 Terahertz imaging

Terahertz imaging is an innovative technology of imaging which is exploited in several applications, such as medical diagnosis, quality control, security, and biological and chemical identification [63, 83, 82, 54, 112, 85, 113, 101, 6, 4, 5, 3]. Terahertz images can be measured by acquisition of sequences of Terahertz pulses, called time series or signals, reflected from or transmitted through a sample. Each time series is assumed univariate and can be represented by several temporal bands (e.g. 1000 bands) behind every pixel. The huge amount of raw bands can be a hurdle to analyze this type of data. Moreover, some bands can be noisy, redundant or uninformative for further processing. The bands, called features, are used for the segmentation of Terahertz images. In the most related works, classification of bands is used for Terahertz image segmentation.

In the state of the art, Terahertz image processing can be used with a feature space constituted by all the bands or only a single band. The selection of a single band can be fixed priori from the Terahertz image and its processing does not lead to satisfactory results. The reader can find more about details in [14, 22]. For multiband space, the pixels of the THz image are considered as vectors and the processing is equivalent to implement classification algorithms [109, 14, 13]. The vectors can be represented by many bands, such as the entire univariate time series of the Terahertz image, the entire spectral amplitude, and a set of some selected bands [21, 110, 50, 22, 13]. In [15], authors have proposed a modified $K$-means clustering approach based on ranked set sampling. This approach uses vectors constituted by both univariate time series and its spectral transformation to analyze Terahertz images. It is essentially less sensitive to the initialization of the centers. Some authors have proposed to automatically reduce the feature space into the clustering process. Recent research is proposed in [13] which presents a feature selection strategy and a random sampling design in $K$-means clustering for THz image segmentation. Automatic estimation of the random sample size and the selected feature number are also proposed. Other authors found useful to extract lower dimensional features before classifying the univariate time series. Some useful features are extracted by using principal component analysis (PCA) [85, 113, 14, 31], decision tree [50], autoregressive (AR) model and autoregressive moving average model (ARMA) [109, 14]. In [109], the parameters of AR and ARMA models have been combined together as a small feature space. Then, Mahalanobis distance classifier has been used to assist biomedical diagnosis and mail/packaging inspection. In [14], both AR parameters and PCA features form a vector characterizing each pixel of THz image. $K$-harmonic-means clustering technique was then used on the extracted features to segment THz images.

Among these techniques, AR models are well adapted to characterize THz time series before the clustering process. In the following section, we will present how to model the time series by using univariate AR models and then reformulate it to segment the Terahertz images.

## 2.2 Autoregressive Modeling

Autoregressive (AR) models have been largely used in several pattern recognition applications [40, 105]. Univariate AR models consist to model the current value of the time series variable as a weighted linear sum of its previous values plus an error, considered as a centered Gaussian random variable of variance $\sigma_t^2$. Because the time series is stationary, the $\sigma$ does not depend on $t$. The order of the model is the number of preceding observations used, and the weights (also called coefficients) characterize the time series. We consider an ergodic discrete-time random process $X = \{X_1, \ldots, X_T\}$, its realizations noted in vector form $x = (x_1, \ldots, x_T)'$, and a model order $P$. The AR($P$) model predicts the next value $x_t$ in the time series as a linear combination of the $P$ previous values. The AR coefficients $w = (w_1, \ldots, w_P)'$ will be determined by fitting the model to the training time series data. This can be done by minimizing an error function that measures the misfitting between the predicted model, for any given values of $w$, and the training time series data points. Let us consider $\phi_t = (x_{t-1}, x_{t-2}, \ldots, x_{t-P})'$ which represents a vector of the $P$ previous realizations of $x_t$. The most used error function is given by the sum of the squares of the errors between the target values $x_t$ and the corresponding predicted values $\phi_t' w$, so that we minimize

$$J_{AR} = \sum_{t=P+1}^{T} (x_t - \phi_t' w)^2 \tag{4.1}$$

This fitting problem is solved by choosing the value of $w$ for which the error function is as small as possible. Because the error function is a quadratic function of the coefficients $w$, its derivatives with respect to the coefficients will be linear in the elements of $w$, and so the minimization of the error function with respect to $w$ has an unique solution which can be found in closed form. We can deduce then the least squares solution for the autoregressive coefficients as follows

$$w = \left( \sum_{t=P+1}^{T} \phi_t \phi_t^{'} \right)^{-1} \left( \sum_{t=P+1}^{T} \phi_t x_t \right) \tag{4.2}$$

By using the estimated autoregressive coefficients $w$, the noise variance $\sigma^2$ is given by

$$\sigma^2 = \frac{1}{T-P} J_{AR} \tag{4.3}$$

In fact, each univariate time series $x$ was assumed originally generated by an AR model. In this section, we have seen how to recover the original model by estimating their coefficients $w$. The reader can find more about details in [29]. Let us recall that each pixel of the Terahertz image is generated from an autoregressive model. In the next section, we will formulate the segmentation of Terahertz images by using the clustering of AR models.

## 3 $K$-Autoregressive Clustering

Standard clustering algorithms, such as $K$-means techniques [77, 67, 73, 57, 35], were largely used for Terahertz image analysis [85, 21, 22, 13, 15]. These techniques were generally involved with clusters defined by measures of the central tendency, called arithmetic means or centers, and pixels described by the whole feature space. These pixels are depicted by univariate time series generated by serial correlated variables and generally classified by using Euclidian distance measure. However, these are not desirable in Terahertz imaging where pixels are represented by a huge number of raw bands. The relevance problem of these bands can be a hurdle to analyze this type of images and the correlation between variables was not used in the foremost time series clustering techniques which decreases its performances. Moreover, arithmetic means are not suitable to represent time series clusters and other statistical methods more adapted to time series can improve the clustering process [96]. The regions in THz images are shown as clusters of time series [13, 108]. We believe that autoregressive modeling leads to model each cluster of univariate time series. In this section, we propose a novel clustering approach called $K$-Autoregressive ($K$-AR) model which consists to classify the set of THz time series into a fixed number of clusters, each one is represented by its predictive prototype. We assume that each pixel of the Terahertz image is a stationary time series generated from an AR model where the order is unknown. The $K$-AR approach consists to classify the AR models and minimize a suitable objective function for recovering the original $K$ autoregressive models describing each cluster of time series. Let us consider $N$ discrete-time random process $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_N\}$. Since each $\mathcal{X}_n = \{X_{n1}, \ldots, X_{nT}\}$ is an ergodic process, its realizations correspond to the $n^{th}$ pixel and noted in vector form $x_n = (x_{n1}, \ldots, x_{nT})^{'}$. $T$ is the number of realizations, $P$ is the AR order, the weights $w_k = (w_{k1}, \ldots, w_{kP})^{'}$ are the AR coefficients which characterize the $k^{th}$ time series cluster and $\phi_{nt} = (x_{n(t-1)}, x_{n(t-2)}, \ldots, x_{n(t-P)})^{'}$ is a vector of the $P$ previous realizations of $x_{nt}$. For each cluster, the fitting error of an AR($P$) is $\sum_{x_n \in C_k} \sum_{t=P+1}^{T} (x_{nt} - \phi_{nt}^{'} w_k)^2$, where $C_k$ is the set of time series of the $k^{th}$ cluster. For all clusters, this error is equal to the sum of $K$ AR($P$) fitting errors. We need to find coefficients $W = (w_1, ..., w_K)$, which minimize the above error, for all pixels:

$$J_{KAR} = \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{t=P+1}^{T} u_{nk} \left( x_{nt} - \phi_{nt}^{'} w_k \right)^2 \tag{4.4}$$

Equation (4.4) is solved by choosing the values of $w_k$ for which $J_{KAR}$ is as small as possible. Because the objective function is a quadratic function of the coefficients $w_k$, its derivatives with respect to the coefficients will be linear in the elements of $w_k$, and so the minimization of the objective function $J_{KAR}$ in equation (4.4) with respect to $w_k$ gives the following expression

$$\sum_{n=1}^{N} \sum_{t=P+1}^{T} u_{nk} \phi_{nt} x_{nt} - \sum_{n=1}^{N} \sum_{t=P+1}^{T} u_{nk} \phi_{nt} \phi_{nt}^{'} w_k = 0 \tag{4.5}$$

Then, we can deduce the expression of the AR weights as following

$$w_k = \left(\sum_{n=1}^{N}\sum_{t=P+1}^{T} u_{nk}\phi_{nt}\phi_{nt}^{'}\right)^{-1}\left(\sum_{n=1}^{N}\sum_{t=P+1}^{T} u_{nk}\phi_{nt}x_{nt}\right). \tag{4.6}$$

This equation shows that AR weights are estimated by using the structure of time series clusters represented by $u_{nk}$ and the empirical correlated variables represented by both terms $\phi_{nt}\phi_{nt}^{'}$ and $\phi_{nt}x_{nt}$. The time series are assigned to their closest cluster by computing the membership degrees $u_{nk}$. The values of the membership degrees must verify the constraints $\{u_{nk} \mid u_{nk} \in \{0,1\}$ and $\sum_{k=1}^{K} u_{nk} = 1\}$. The necessary condition for minimizing the objective function $J_{KAR}$ gives the following expression

$$u_{nk} = \begin{cases} 1, & \text{if } k = \arg\min_l\{\mathcal{D}_{nl}\} \\ 0, & \text{otherwise} \end{cases} \tag{4.7}$$

where $\mathcal{D}_{nk} = \sum_{t=P+1}^{T}(x_{nt} - \phi_{nt}^{'}w_k)^2$ represents the sum of the squares of the errors between the target values $x_{nt}$ and the corresponding predicted values $\phi_{nt}^{'}w_k$. We assume that the number $K$ of clusters is known. For clustering, an AR model is assigned with the cluster $k$ if $u_{nk} = 1$. The $K$-AR consists to classify the set of time series data into $K$ clusters; each cluster is represented by one AR weights $w_k$, where $k = 1, \ldots, K$. Our approach integrates correlated variables information into the clustering process. The learning process is then done by iterating between two steps, updating the AR weights of the clusters and the membership of the time series, until convergence, i.e. when the value of the objective function is minimal. Let us consider a parameter $P$ much less than $T$. The $K$-AR algorithm can be summarized as follows:

```
K-AR algorithm
1. Initialize $w_k$ by random values
2. Do
       Update membership degrees $u_{nk}$ using Eq. (4.7)
       Update AR weights $w_k$ using Eq. (4.6)
   Until $|J_{KAR}^{(t)} - J_{KAR}^{(t-1)}|$ <threshold
3. Return $U$.
```

The pixels represented by the time series are then assigned to the clusters having the best AR model fitting. The resulted clusters are defined by the obtained membership degrees of the time series. The computational complexity of $K$-AR algorithm is $O(KNTP^2)$ for one iteration, where $N$ is the number of time series, $T$ is the number of observations in the time series, $K$ is the number of clusters and $P$ is the AR model order. This complexity is linear for parameters $K$, $N$, $T$ and quadratic for parameter $P$.

## 4 Parameter Selection

This section deals with the selection of two fundamental parameters for our approach, the number $K$ of clusters and the order $P$ of AR models. AR model order is generally selected using information criteria in which the goal is to achieve the best tradeoff between model fitting and model complexity [7, 98, 49, 30]. Considering one stationary time series, the parameter $\sigma^2$, as defined in equation (4.3), is considered as the residual variance of the error function between the time series observations and the corresponding predicted values. Among different AR model orders, the best models fitting are that which have the minimal error variances. Generalized information criterion (GIC) [30] is an objective measure largely used for model selection.

Considering $N$ stationary univariate time series $x_n = (x_{n1}, \dots, x_{nT})'$, where $n = 1, \dots, N$, the objective function $J_{KAR}$ defined in equation 4.4 is considered as the sum of errors of $T$ realizations for the $N$ time series regrouped into $K$ clusters. By analogy with one univariate time series case (see equation 4.3), the global error variance corresponding to $N$ time series is given by

$$\Sigma^2_{K,P} = \frac{1}{N(T-P)} J_{KAR} \tag{4.8}$$

In this section, we propose a modified generalized information criterion (MGIC) well adapted to a set of time series which balances between model fitting and model complexity. MGIC introduces a penalty term for the number of parameters in the model. The proposed criterion consists to estimate the best $K$-AR model parameters, i.e. the model order $P$ and the clusters number $K$. It is defined as follows:

$$MGIC(P,K,\alpha) = -\log \Sigma^2_{P,K} - \alpha\left(\frac{P \times K + 1}{T + N}\right), \tag{4.9}$$

where $P \times K + 1$ represents the number of parameters to be estimated, $N + T$ represents the sum of data points number in $X$ and the features number in each $X_n$, and $\alpha$ is a positive real penalty factor. This parameter is set by trial and error procedure. Maximizing MGIC criterion return to minimizing the error of the noise and the number of the used parameters. To select a model order and a cluster number, MGIC($P$,$K$,$\alpha$) criterion is determined for all $P$ between 1 and a maximum candidate order $P_{max}$ and for all $K$ between 2 and a maximum candidate number $K_{max}$. The order and the clusters number with the maximal value of the criterion are selected.

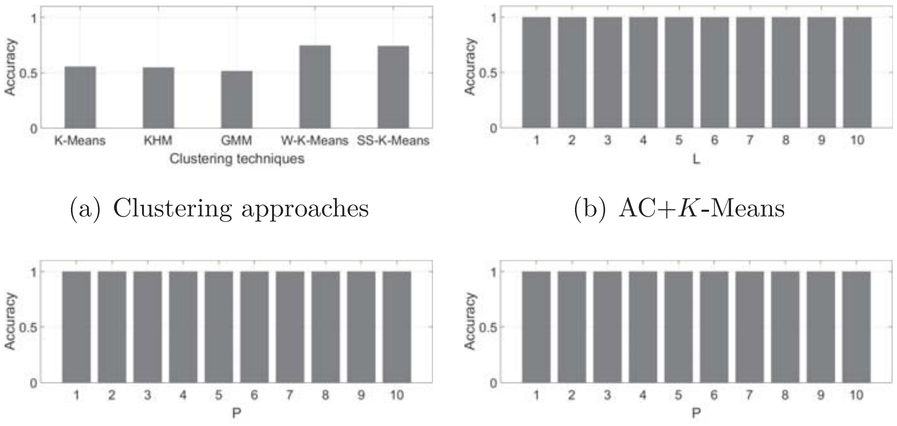
# 5 Experimental results

In order to better understand the properties of our approach, artificial data sets are used firstly to validate the clustering algorithms. Afterward, THz images are used to examine the different algorithms. In this section, we experimentally and empirically show that our clustering algorithm outperforms other algorithms such as standard $K$-means [77], Gaussian mixture model (GMM) [41], W-$K$-means [64], $K$-harmonic-means (KHM) [111], SS-$K$-means [13], $K$-means based autocorrelation functions (AC+$K$-means) [48] and $K$-means based autoregressive coefficients (AR+$K$-means) [48]. AC+$K$-means and AR+$K$-means are used in these experiments with hard decision rule. The details about $K$-AR algorithm implementation are presented in our website [1].

## 5.1 Experiments on artificial data sets

Artificial data sets are often used to validate the clustering algorithms. In this experiment, we used three artificial data sets with various cluster number to verify the performances of the clustering algorithms. The three data sets $D_P$ were constituted by

[1] https://ayechwalid.wixsite.com/ayechwalid



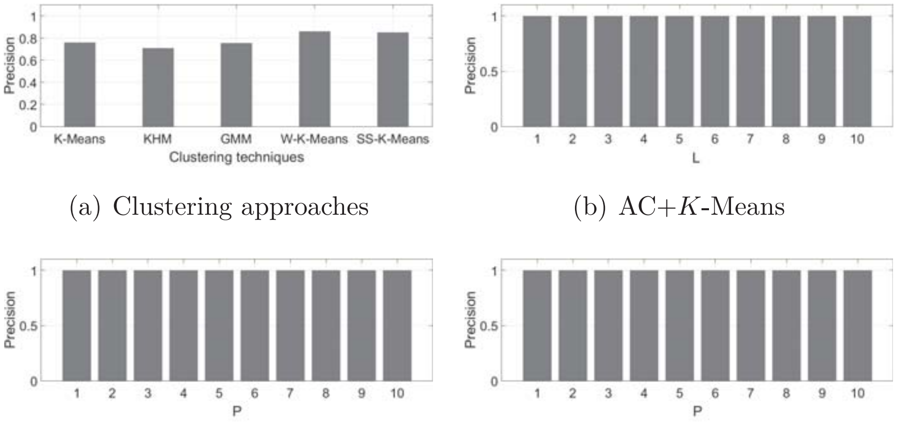
(a) Clustering approaches

(b) AC+$K$-Means

(c) AR+$K$-Means

(d) $K$-AR

Figure 4.1 – Accuracy of clustering algorithms on dataset $D_2$. The parameter $K$ is fixed to 2.



(a) Clustering approaches

(b) AC+$K$-Means

(c) AR+$K$-Means

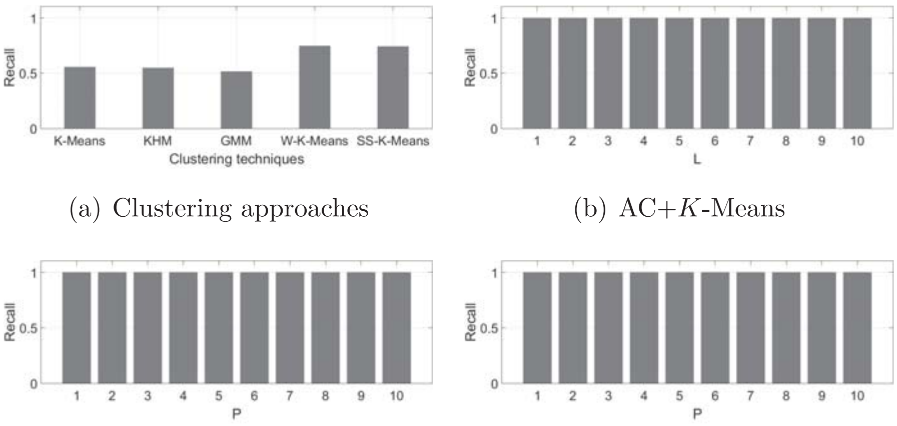
(d) $K$-AR

Figure 4.2 – Precision of clustering algorithms on dataset $D_2$. The parameter $K$ is fixed to 2.

$K$ clusters of time series originated from univariate AR($P$) process using the following equation

$$x_{nt} = \phi_{nt}^{'} w_n + \varepsilon_{nt} \tag{4.10}$$

where $x_{nt}$ is the $t^{th}$ observation of the $n^{th}$ time series, $w_n$ is a column vector of $P$ AR weights for the $n^{th}$ time series, $\phi_{nt} = (x_{n(t-1)}, x_{n(t-2)}, \ldots, x_{n(t-P)})'$ are the $P$ previous time series observations of $x_{nt}$ and $\varepsilon_{nt}$ is additive Gaussian noise with zero mean



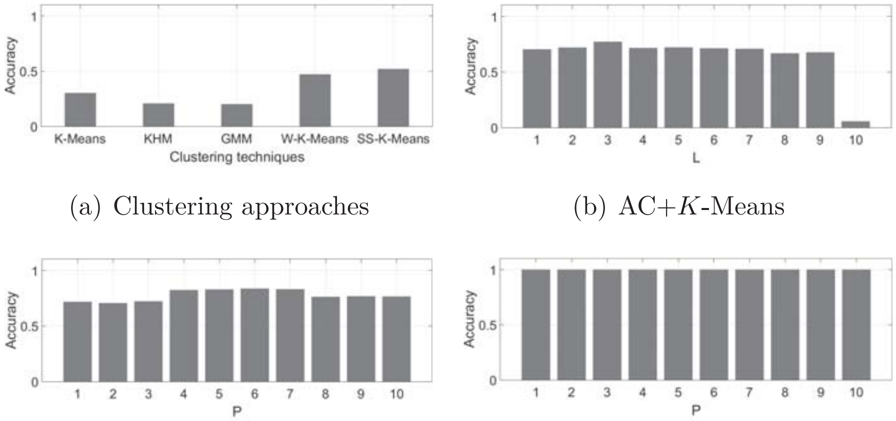
(a) Clustering approaches

(b) AC+$K$-Means

(c) AR+$K$-Means

(d) $K$-AR

Figure 4.3 – Recall of clustering algorithms on dataset $D_2$. The parameter $K$ is fixed to 2.



(a) Clustering approaches

(b) AC+$K$-Means

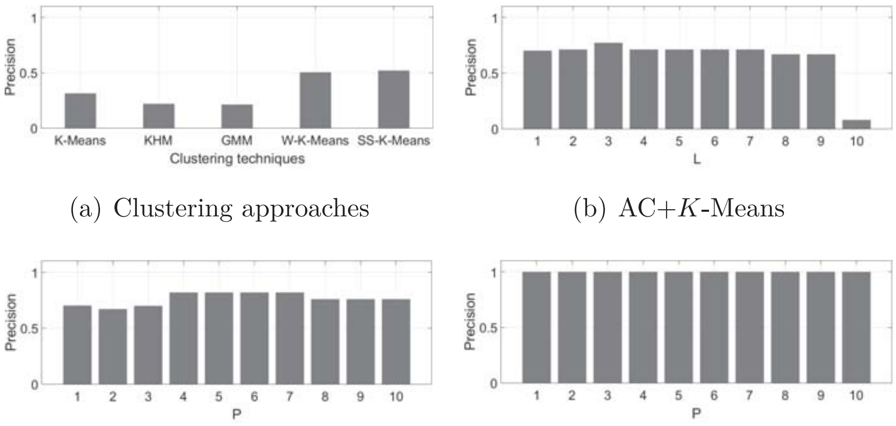
(c) AR+$K$-Means

(d) $K$-AR

Figure 4.4 – Accuracy of clustering algorithms on dataset $D_3$. The parameter $K$ is fixed to 5.

and $\sigma_n^2$ variance. For the different time series, the variances of the Gaussian noise are randomly chosen between 0.001 and 0.1.

In these experiments, data set $D_2$ was constituted by 400 stationary univariate time series distributed into two clusters originated from univariate AR(2) process. Each time series is characterized by 500 bands. The AR weights $w_n$ are equal to $(1.8, -.9)'$ and $(.5, -.3)'$ respectively for time series of cluster 1 and cluster 2. Data



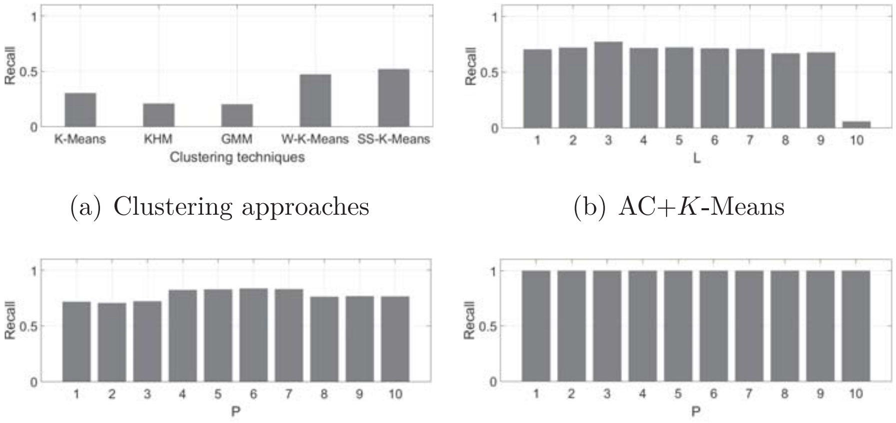
(a) Clustering approaches

(b) AC+$K$-Means

(c) AR+$K$-Means

(d) $K$-AR

Figure 4.5 – Precision of clustering algorithms on dataset $D_3$. The parameter $K$ is fixed to 5.



(a) Clustering approaches

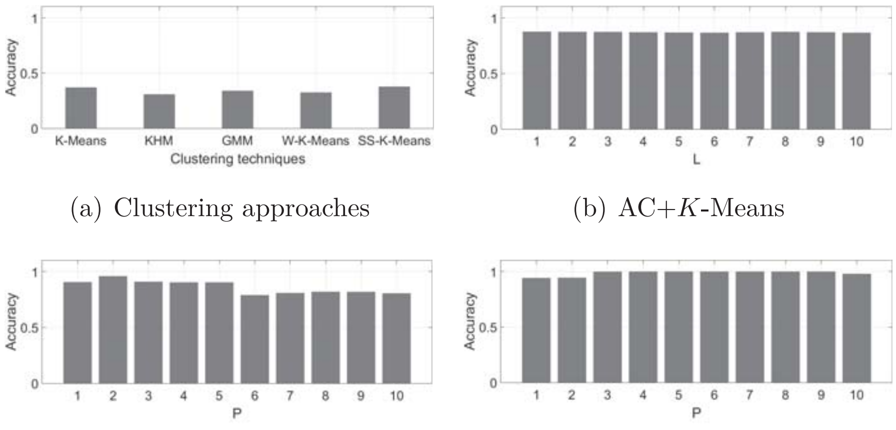
(b) AC+$K$-Means

(c) AR+$K$-Means

(d) $K$-AR

Figure 4.6 – Recall of clustering algorithms on dataset $D_3$. The parameter $K$ is fixed to 5.

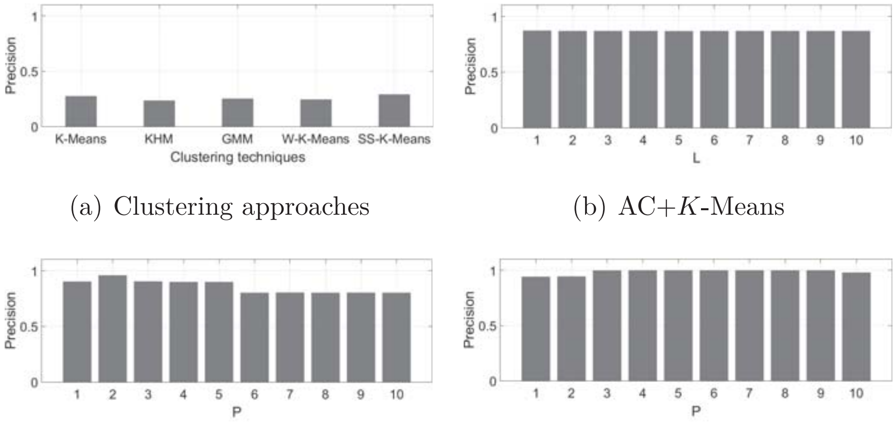
set $D_3$ was constituted by 300 stationary univariate time series distributed into five clusters originated from univariate AR(3) process. Each time series is characterized by 1000 bands. The AR weights $w_n$ are equal to $(-.6, -.1, -.1)'$, $(.005, 0, -.001)'$, $(.1, -.5, .5)'$, $(.5, .2, .2)'$ and $(1, -.6, .2)'$ respectively for cluster 1, cluster 2, cluster 3, cluster 4 and cluster 5. Data set $D_5$ was constituted by 500 stationary univariate time series distributed into three clusters originated from AR(5) process. Each uni-



(a) Clustering approaches (b) AC+$K$-Means

(c) AR+$K$-Means (d) $K$-AR

Figure 4.7 – Accuracy of clustering algorithms on dataset $D_5$. The parameter $K$ is fixed to 3.

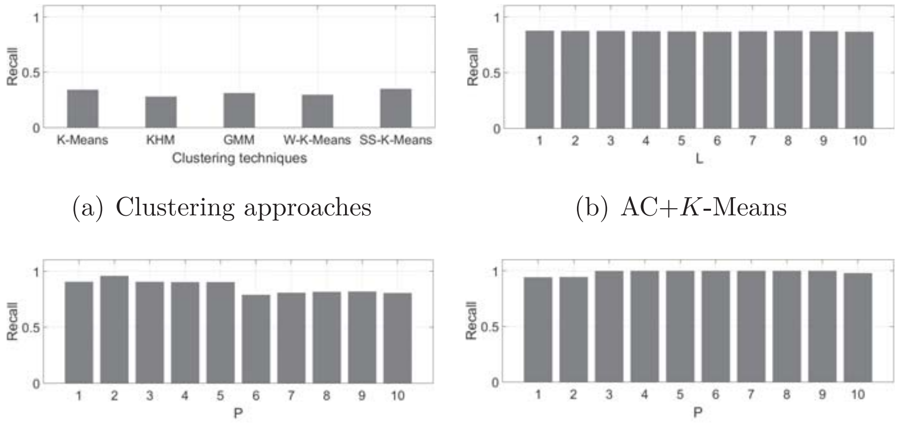


(a) Clustering approaches (b) AC+$K$-Means

(c) AR+$K$-Means (d) $K$-AR

Figure 4.8 – Precision of clustering algorithms on dataset $D_5$. The parameter $K$ is fixed to 3.

variate time series is characterized by 500 bands. The AR weights $w_n$ are equal to $(-.3, -.1, -.1, .1, -.1)'$, $(.005, 0, -.001, -.005, .005)'$ and $(.3, .1, .2, -.2, -.2)'$ respectively for cluster 1, cluster 2 and cluster 3. The different clustering techniques were statistically compared in terms of clustering performance. We used accuracy, precision and recall to evaluate the results [93]. The accuracy is the proportion of time series correctly classified. The precision is computed as the fraction of classified time



(a) Clustering approaches

(b) AC+$K$-Means

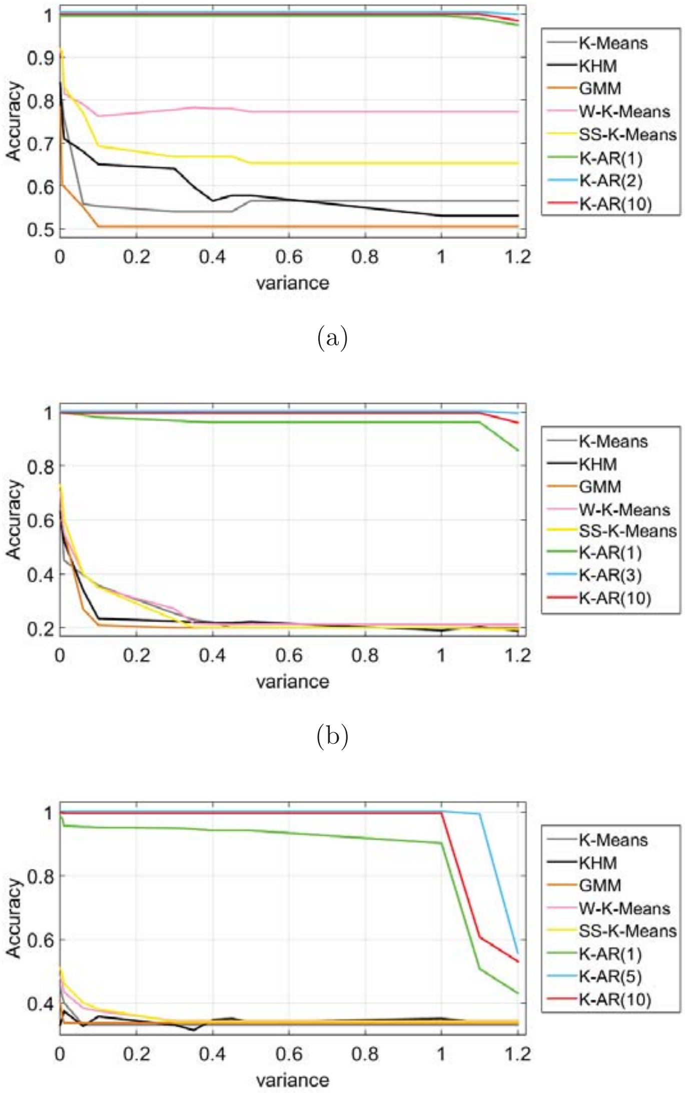
(c) AR+$K$-Means

(d) $K$-AR

Figure 4.9 – Recall of clustering algorithms on dataset $D_5$. The parameter $K$ is fixed to 3.

series which belong to the relevant class. The recall is computed as the fraction of the relevant time series which are correctly classified.

This section consists to evaluate the performances of the clustering techniques $K$-means, KHM, GMM, W-$K$-means, SS-$K$-means, AC+$K$-means for different values of the time lag parameter $L$, AR+$K$-means and $K$-AR for different values of the AR parameter $P$. SS-$K$-means was used with parameters $a = 2$, $b = 2$ and $c = 1$. Figures from 4.1 to 4.9 comprise the average evaluation measures of five different runs of the three artificial data sets. For data set $D_2$, the clustering accuracy is less than 0.6 for $K$-means, GMM and KHM and around 0.75 for W-$K$-means and SS-$K$-means. While, it is equal to 1.0 for AC+$K$-means with $L$ from 1 to 10, AR+$K$-means and $K$-AR clustering with $P$ from 1 to 10. The precision measure is between 0.7 and 0.78 for $K$-means, GMM and KHM, around 0.82 for W-$K$-means and SS-$K$-means, and equal to 1.0 for AC+$K$-means, AR+$K$-means and $K$-AR clustering for different values of $L$ and $P$. Also, recall measure does not surpass 0.6 for $K$-means and KHM, around 0.5 for GMM, 0.75 for W-$K$-means and SS-$K$-means and equal 1.0 for AC+$K$-means, AR+$K$-means and $K$-AR clustering for different values of $L$ and $P$. In figures 4.4, 4.5 and 4.6, clustering accuracy, precision and recall of artificial data set $D_3$ are 0.3 for $K$-means, around 0.23 for GMM and KHM, and does not surpass 0.76 for W-$K$-means and SS-$K$-means. For AC+$K$-means, these measures are between 0.65



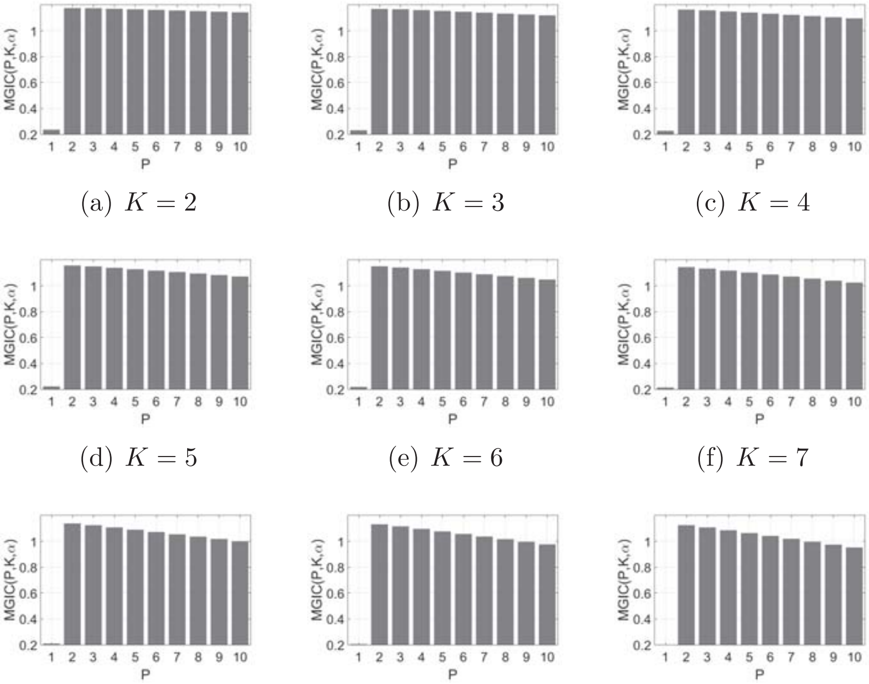
(a)

(b)

(c)

Figure 4.10 – The variation of clustering accuracy in terms of error variance $\sigma_n^2$ for data sets $D_2$ (a), $D_3$ (b) and $D_5$ (c). Each data set was generated for different values of $\sigma_n^2$ from $10^{-4}$ to 1.2. The parameter $K$ is fixed to 2, 5 and 3 for data sets $D_2$, $D_3$ and $D_5$.

and 0.75 for $L$ from 1 to 9, and decrease to 0.1 when $L = 10$. For AR+$K$-means, performance measures are around 0.8 for $P$ between 4 and 7, and does not surpass



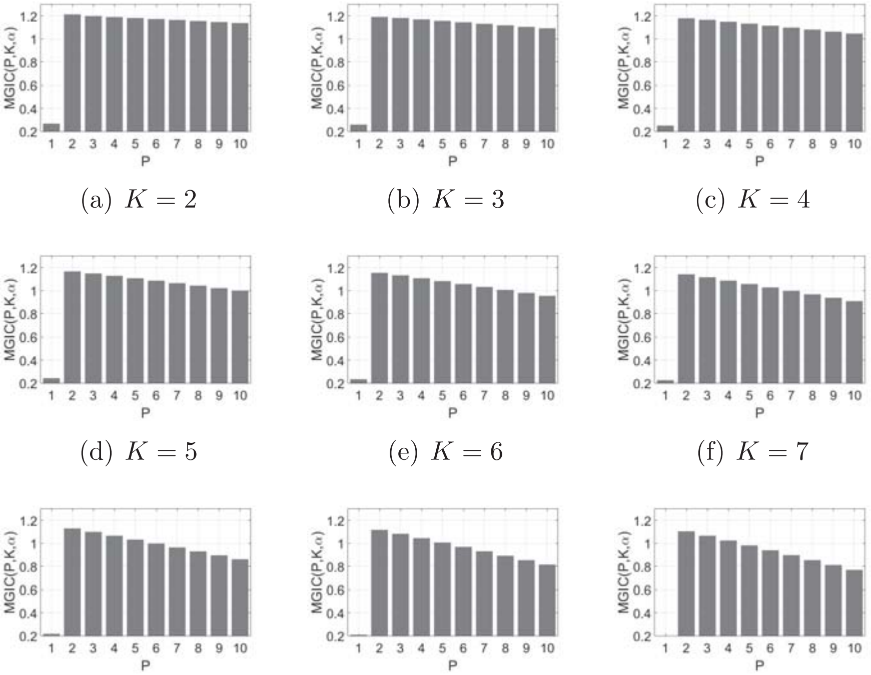
(a) $K = 2$ (b) $K = 3$ (c) $K = 4$

(d) $K = 5$ (e) $K = 6$ (f) $K = 7$

(g) $K = 8$ (h) $K = 9$ (i) $K = 10$

Figure 4.11 – MGIC criterion for data set $D_2$. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ from 2 to 10 (from the left to the right and from the top to the bottom) and $P$ from 1 to 10. The parameter $\alpha$ is fixed to 2.

0.75 for $P < 4$ and $P > 7$. While, they are around to 0.93 for $K$-AR with $P = 1$ and 0.97 for $K$-AR with $P > 1$. In figures 4.7, 4.8 and 4.9, clustering accuracy, precision and recall of artificial data set $D_5$ are between 0.24 and 0.36 for $K$-means, GMM, KHM, W-$K$-means and SS-$K$-means, around 0.85 for AC+$K$-means and around 0.9 for AR+$K$-means with $P$ between 1 and 5 and around 0.77 when $P > 5$. These measures are around 0.9 for $K$-AR with $P = 1$ and $P = 2$ and become 1.0 for $K$-AR when $P > 3$. The obtained statistics show the high performances of $K$-AR compared to the other approaches. Figures 4.10 (a), (b) and (c) present the variation of clustering accuracy in terms of $\sigma_n^2$ for data sets $D_2$, $D_3$ and $D_5$. Each data set was generated for different values of $\sigma_n^2$ (from $10^{-4}$ to 1.2) and then classified using different clustering techniques. These figures show that $K$-AR approach outperforms



(a) $K = 2$

(b) $K = 3$

(c) $K = 4$

(d) $K = 5$

(e) $K = 6$

(f) $K = 7$

(g) $K = 8$

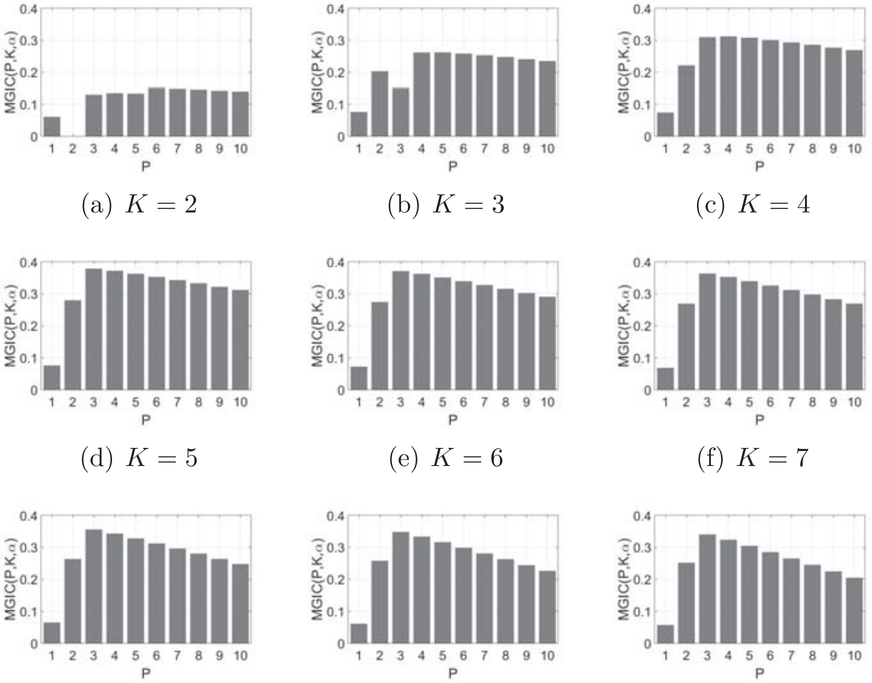
(h) $K = 9$

(i) $K = 10$

Figure 4.12 – MGIC criterion for data set $D_2$. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ from 2 to 10 (from the left to the right and from the top to the bottom) and $P$ from 1 to 10. The parameter $\alpha$ is fixed to $2(\log(\log(T + N)))$.

the other approaches especially when the error variance is inferior to 1. However, as foremost clustering techniques, $K$-AR is sensitive to initial starting parameters. For twenty different runs, the $K$-AR performances can change after each run. For dataset $D_3$, the accuracy measure of $K$-AR clustering with $P = 3$ are equal to 1.0 for 60% of tests and between 0.7 and 0.75 for the rest of tests. For dataset $D_5$, the accuracy of $K$-AR with $P = 5$ are equal to 1.0 for 90% of tests and near to 0.55 for the rest of tests. While for dataset $D_2$, $K$-AR with $P = 2$ produces a high accuracy measure equal to 1.0 for 100% of tests. These statistics show the sensitivity of $K$-AR to initial values of AR weights and this degree of sensitivity can depend on the structure of the tested datasets.

Clusters number $K$ and AR model order $P$ are selected by using MGIC($P$,$K$,$\alpha$)



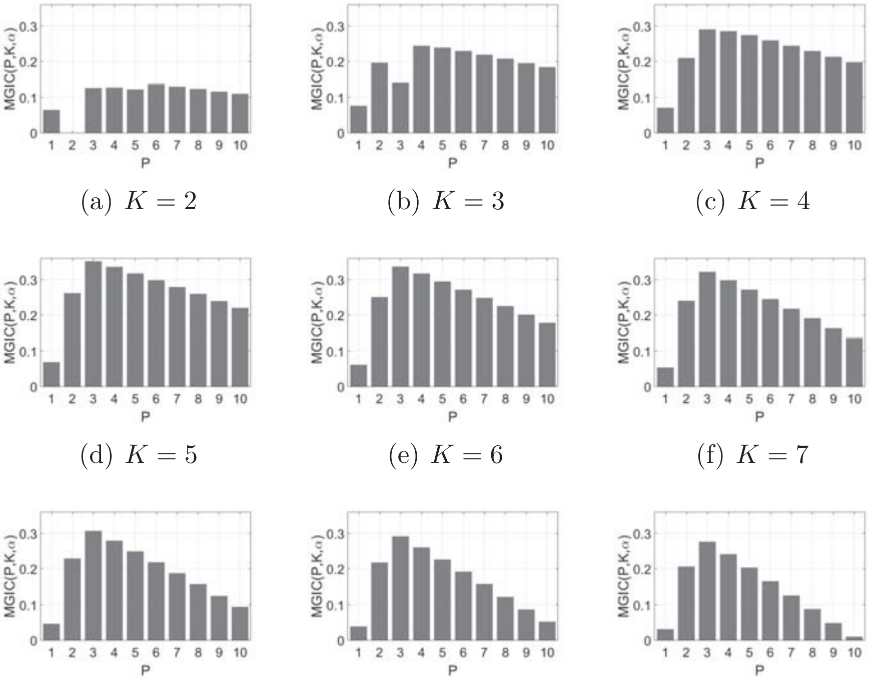
(a) $K = 2$ (b) $K = 3$ (c) $K = 4$

(d) $K = 5$ (e) $K = 6$ (f) $K = 7$

(g) $K = 8$ (h) $K = 9$ (i) $K = 10$

Figure 4.13 – MGIC criterion for data set $D_3$. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ from 2 to 10 (from the left to the right and from the top to the bottom) and $P$ from 1 to 10. The parameter $\alpha$ is fixed to 2.

criterion. In this paper, we study the parameter selection of our clustering approach for two different values of $\alpha$ ($\alpha = 2$ and $\alpha = 2(log(log(T + N))))$. MGIC with both values of $\alpha$ can be considered as modified versions of Akaike information criterion [7] and Hannan-Quinn information criterion [49], respectively. Figures 4.11, 4.12, 4.13, 4.14, 4.15, and 4.16 show the corresponding MGIC measures of $K$-AR($P$) as functions of $P$ for nine different values of $K$. These figures show that MGIC values give a correct AR model order estimation for different cluster number. The highest MGIC measure corresponds to $K$ equal to 2, 5 and 3 and $P$ equal to 2, 3 and 5, respectively for data sets $D_2$, $D_3$ and $D_5$. These figures show that $P$ and $K$ are well estimated by MGIC criterion and confirm the results found previously in figures from 4.1 to 4.9. The estimated values of $P$ and $K$ confirm the best tradeoff between the clustering



(a) $K = 2$ (b) $K = 3$ (c) $K = 4$

(d) $K = 5$ (e) $K = 6$ (f) $K = 7$

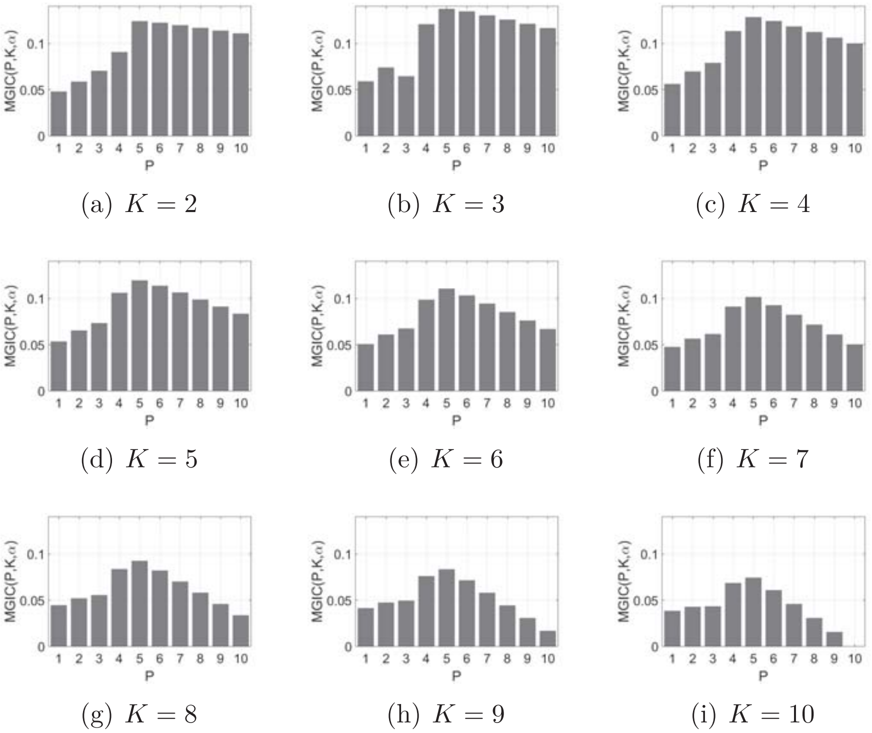
(g) $K = 8$ (h) $K = 9$ (i) $K = 10$

Figure 4.14 – MGIC criterion for data set $D_3$. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ from 2 to 10 (from the left to the right and from the top to the bottom) and $P$ from 1 to 10. The parameter $\alpha$ is fixed to $2(\log(\log(T+N)))$.

performances and the low computational cost.

## 5.2 Experiments on Terahertz images segmentation

In this section, $K$-AR, $K$-means, W-$K$-means, KHM, GMM, SS-$K$-means, AC+$K$-means and AR+$K$-means are tested on cork, chemical and moth THz images. The cork image is acquired from the department of physics, New Jersey Institute of Technology, while the chemical and the moth images are acquired from the company Zomega Terahertz Corporation. Pixels of THz images are formed respectively by 1024, 1052 and 894 bands in the time domain. Since Terahertz images cannot be visualized (hundreds or thousands of bands), we present in figure 4.17 (a), figures 4.18 (a) and (b) and figure 4.19 (a) the objects acquired in the visible light for the



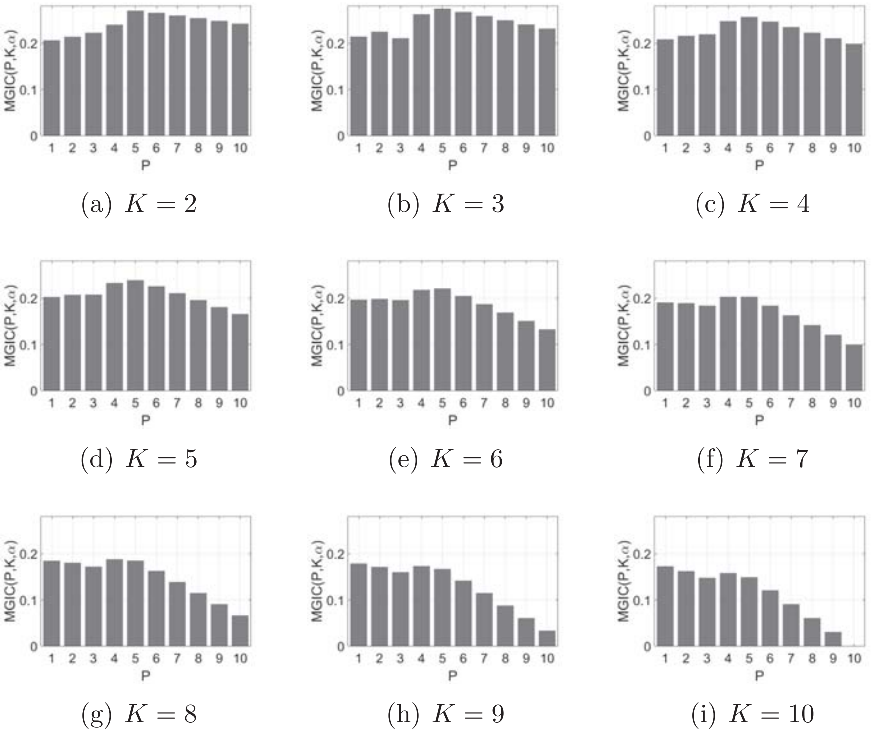
(a) $K = 2$ (b) $K = 3$ (c) $K = 4$

(d) $K = 5$ (e) $K = 6$ (f) $K = 7$

(g) $K = 8$ (h) $K = 9$ (i) $K = 10$

Figure 4.15 – MGIC criterion for data set $D_5$. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ from 2 to 10 (from the left to the right and from the top to the bottom) and $P$ from 1 to 10. The parameter $\alpha$ is fixed to 2.

validation. The ground truth of the chemical image, the $195^{th}$ band of the cork THz image and the $570^{th}$ band of the moth THz image are shown in the right of the same figures. The chemical image comprises four compounds, L-Valine (0.200g), L-Tryptophan (0.100g), L-Tryptophan (0.200g) and Proline (0.200g), distributed into four false colored regions, whereas, the second THz image shows cork matter with some cork grains as well as some voids, defects and cracks, and the moth THz image mainly comprises a body and two wings. The weights of $K$-AR approach and the centers of W-$K$-means, SS-$K$-means, $K$-means, KHM, GMM, AC+$K$-means and AR+$K$-means were initialized by random values. SS-$K$-means was used with parameters $a = 2$, $b = 2$ and three different values of parameter $c$. The segmentation of chemical, cork and moth images was employed respectively with 4, 5 and 5 clusters.



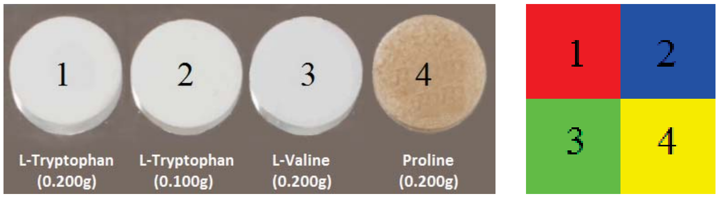
(a) $K = 2$ (b) $K = 3$ (c) $K = 4$

(d) $K = 5$ (e) $K = 6$ (f) $K = 7$

(g) $K = 8$ (h) $K = 9$ (i) $K = 10$

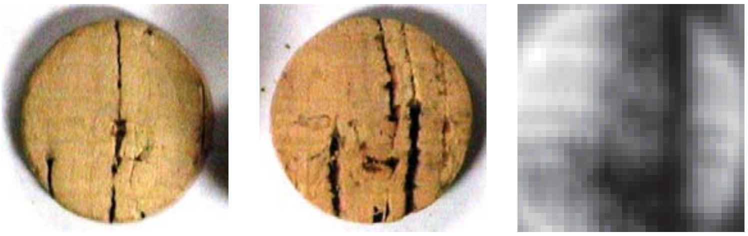
Figure 4.16 – MGIC criterion for data set $D_5$. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ from 2 to 10 (from the left to the right and from the top to the bottom) and $P$ from 1 to 10. The parameter $\alpha$ is fixed to $2(\log(\log(T + N)))$.

As well as artificial data sets, Terahertz images segmentation are evaluated by using accuracy, precision and recall indices which need external references. Figure 4.20 shows the chemical Terahertz image segmentation for the eight clustering algorithms. In figures 4.20 (a), (b), (c), (d), (e), (g), (h), (i), (j) and (k), $K$-means, KHM, GMM, W-$K$-means, SS-$K$-means for $c = 0.5$ and $c = 2.5$, AC+$K$-means for time lags $L = 5$ and $L = 10$, and AR+$K$-means for $P = 5$ and $P = 10$ produce as output over-segmented images. L-Tryptophan (0.100g) and L-Tryptophan (0.200g) clusters are combined together in the case of $K$-means, KHM and GMM which clearly show their segmentation shortcomings. Also, L-Valine (0.200g) and L-Tryptophan (0.200g) clusters are combined together in the case of SS-$K$-means for $c = 2.5$ and AC+$K$-means for $L = 5$ and $L = 10$. L-Tryptophan (0.200g) cluster is largely affected by noisy



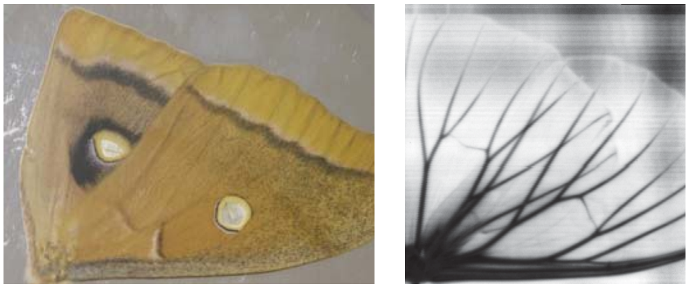
(a) (b)

Figure 4.17 – In the left, an image of four chemical compounds acquired in visible spectrum. In the right, the ground truth of the THz image. The false colors green, blue, red and yellow correspond respectively to the chemical compounds L-Valine (0.200g), L-Tryptophan (0.100g), L-Tryptophan (0.200g) and Proline (0.200g).



(a) (b) (c)

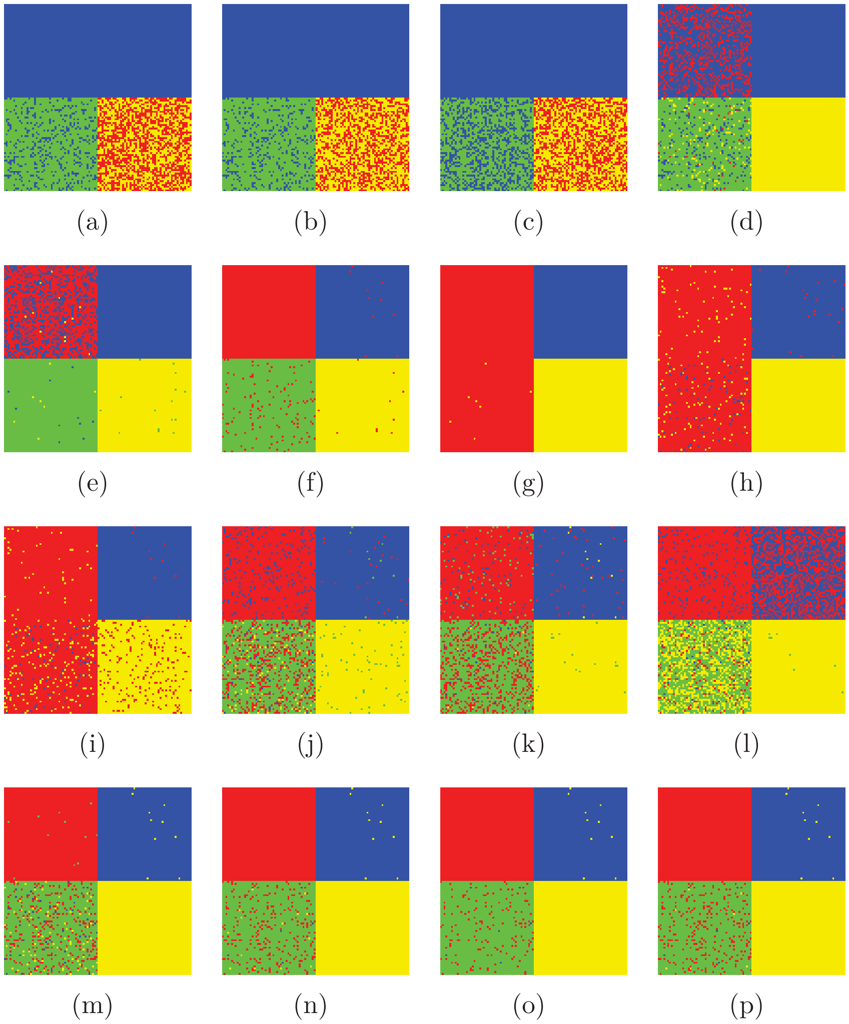
Figure 4.18 – In the left and the middle, an image of a cork acquired in visible spectrum. In the right, the 195$^{th}$ band of the THz image.



(a) (b)

Figure 4.19 – In the left, an image of a moth acquired in visible spectrum. In the right, the 570$^{th}$ band of the THz image.

points in the case of W-$K$-means and SS-$K$-means for $c = 0.5$. For $P = 5$ and $P = 10$, AR+$K$-means produces as output over-segmented regions, the four compounds are



(a) (b) (c) (d)

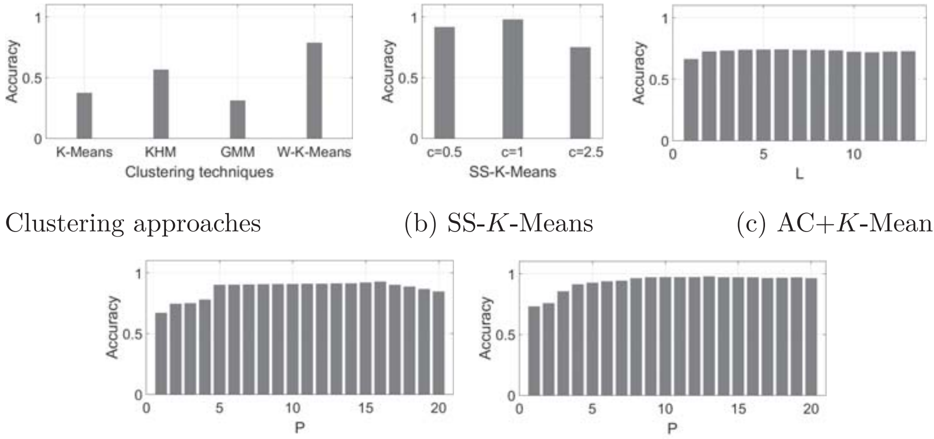
(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

Figure 4.20 – Chemical THz image segmentation for $K$-means (a), KHM (b), GMM (c), W-$K$-means (d), SS-$K$-means for $c = 0.5$ (e), $c = 1$ (f) and $c = 2.5$ (g), AC+$K$-Means for $L = 5$ (h), $L = 10$ (i), AR+$K$-Means for $P = 5$ (j), $P = 10$ (k), and $K$-AR for $P = 2$ (l), $P = 5$ (m), $P = 10$ (n), $P = 13$ (o) and $P = 15$ (p).

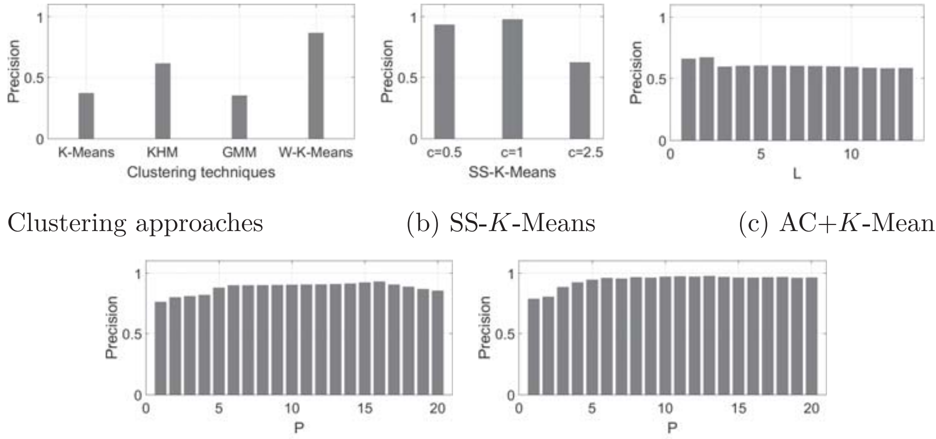
largely affected by noisy points. In figure 4.20 (f), SS-$K$-means for $c = 1$ shows a good segmented regions except some points of the three compounds L-Valine (0.200 g), L-Tryptophan (0.100 g) and Proline (0.200g) are misclassified. Figure 4.20 from (l) to (p) displays the output regions of $K$-AR for different values of $P$ (2, 5, 10,



(a) Clustering approaches (b) SS-$K$-Means (c) AC+$K$-Means

(d) AR+$K$-Means (e) $K$-AR

Figure 4.21 – Accuracy of the clustering algorithms on chemical THz image.

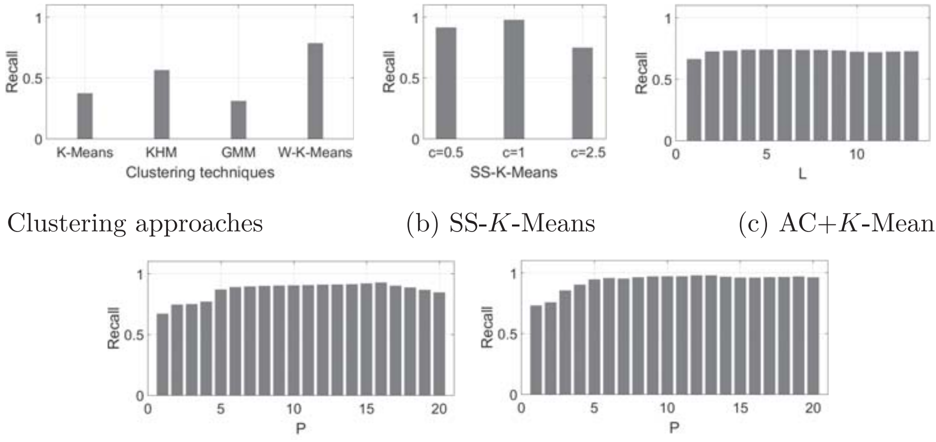


(a) Clustering approaches (b) SS-$K$-Means (c) AC+$K$-Means

(d) AR+$K$-Means (e) $K$-AR

Figure 4.22 – Precision of the clustering algorithms on chemical THz image.

13 and 15). For $P = 2$, the $K$-AR produces as output over-segmented images. The best image segmentations are obtained when $P$ surpasses 2 which appears in figures 4.20 (m), (n), (o) and (p), the four compounds become very well identified, except some points of L-Valine (0.200g) and L-Tryptophan (0.100g) are misclassified. The statistics of the different approaches are shown in figures 4.21, 4.22 and 4.23. The clustering accuracies have not surpassed the 0.4 for $K$-means and GMM, around 0.55 for KHM, 0.76 for W-$K$-means, 0.7 for SS-$K$-Means with $c = 2.5$, surpassed 0.9 for



(a) Clustering approaches (b) SS-$K$-Means (c) AC+$K$-Means

(d) AR+$K$-Means (e) $K$-AR

Figure 4.23 – Recall of the clustering algorithms on chemical THz image.

SS-$K$-Means with $c = 0.5$ and $c = 1$, and between 0.6 and 0.7 for AC+$K$-Means with all values of $L$. For AR+$K$-Means, accuracy measures are between 0.66 and 0.78 for $P$ equal 1, 2, 3 and 4, around 0.9 when $P$ surpass 5 and decrease slightly for $P \geq 17$. While, these measures are around 0.72 for $K$-AR with $P = 1$ and $P = 2$ and increase near to 1.0 when $P > 3$. The precision measures are around 0.37 for $K$-means and GMM, around 0.62 for KHM, around 0.86 for W-$K$-means, surpassed 0.9 for SS-$K$-Means with $c = 0.5$ and $c = 1$, around 0.6 for SS-$K$-Means with $c = 2.5$. Precision measures are around 0.67 for AC+$K$-Means with $L = 1$ and $L = 2$ and around 0.6 for AC+$K$-Means for $L \geq 3$. For AR+$K$-Means, precision measures are between 0.76 and 0.8 for $P$ equal 1, 2, 3 and 4, around 0.9 when $P$ surpass 5 and decrease slightly for $P \geq 17$. However, these measures are around 0.75 for $K$-AR with $P$ equal 1 and 2 and increase and become near to 1.0 when $P$ surpass 3. The recall measures are around 0.4 for $K$-means and GMM, around 0.56 for KHM, between 0.77 and 0.85 for W-$K$-means, surpassed 0.9 for SS-$K$-Means with $c = 0.5$ and $c = 1$, around 0.6 for SS-$K$-Means with $c = 2.5$. Recall measures are between 0.6 and 0.7 for AC+$K$-Means with all values of $L$. For AR+$K$-Means, recall measures are between 0.68 and 0.78 for $P$ equal 1, 2, 3 and 4, around 0.88 when $P$ surpass 5 and decrease slightly for $P \geq 17$. However, these measures are around 0.78 for $K$-AR with $P$ equal 1 and 2 and become near to 1.0 when $P$ surpass 3. The obtained measures confirm the results previously illustrated and show the high performances of our approach.

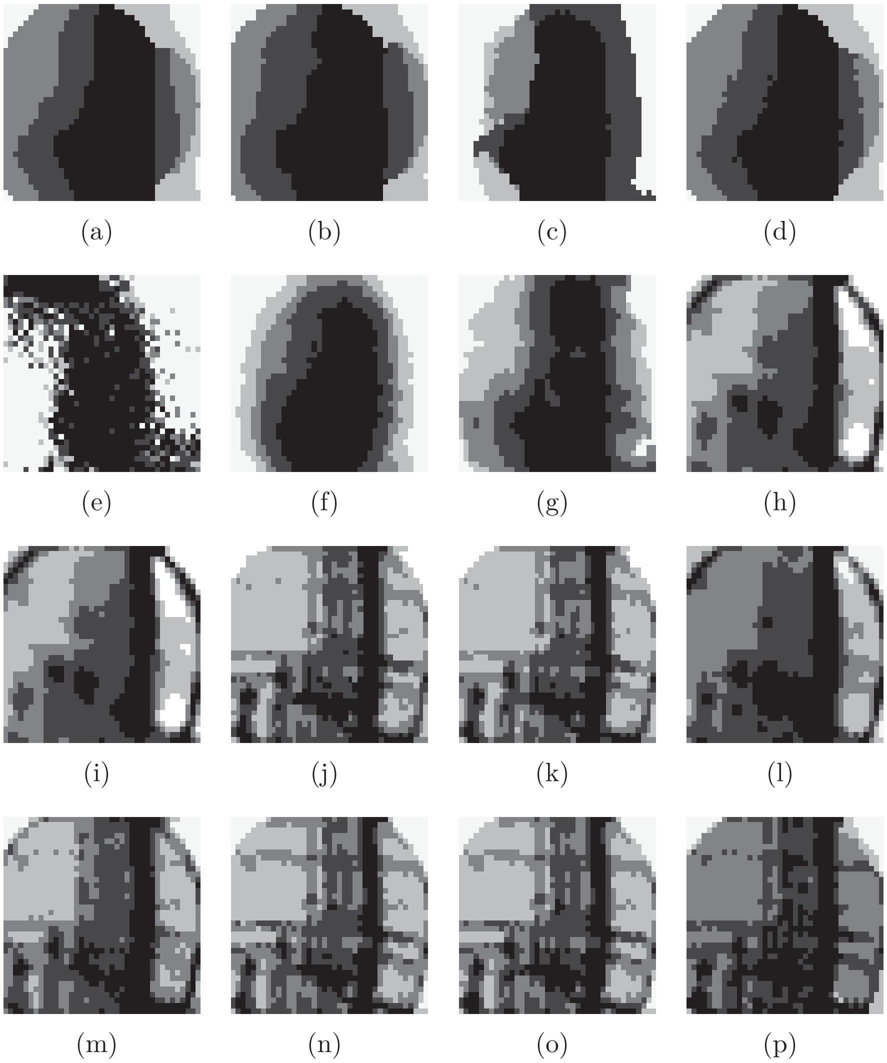


Figure 4.24 – Cork THz image segmentation for $K$-means (a), KHM (b), GMM (c), W-$K$-means (d), SS-$K$-means for $c = 0.5$ (e), $c = 1$ (f) and $c = 2$ (g), AC+$K$-Means for $L = 8$ (h), $L = 10$ (i), AR+$K$-Means for $P = 8$ (j), $P = 10$ (k), and $K$-AR for $P = 1$ (l), $P = 4$ (m), $P = 8$ (n), $P = 10$ (o) and $P = 12$ (p).

Figure 4.24 shows the segmentation outputs of the different clustering algorithms on the cork images. $K$-means, KHM, GMM, W-$K$-means and SS-$K$-means produce a wrongly segmented regions in figures 4.24 (a), (b), (c), (d), (e), (f) and (g). These techniques have not extracted the details inside the cork and clearly illustrate the

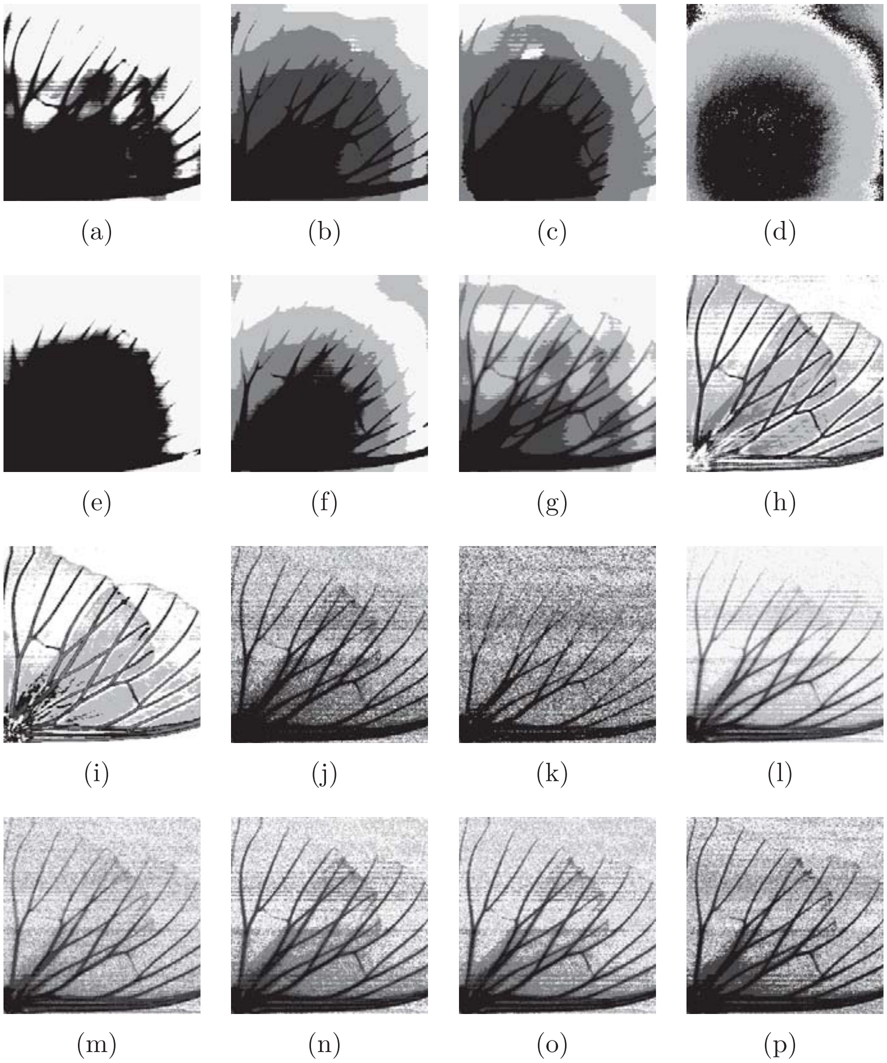
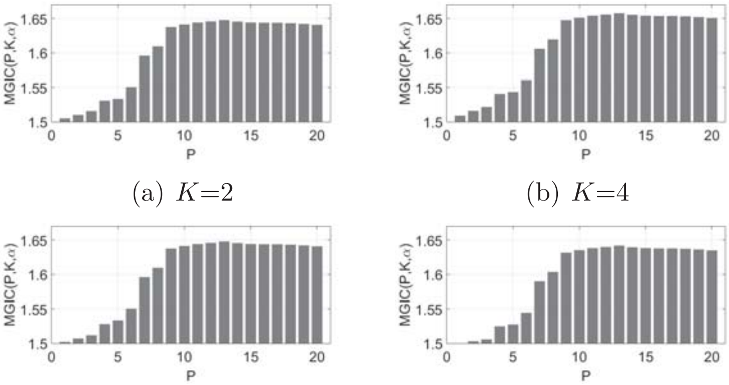


Figure 4.25 – Moth THz image segmentation for $K$-means (a), KHM (b), GMM (c), W-$K$-means (d), SS-$K$-means for $c = 1$ (e), $c = 1.5$ (f) and $c = 2$ (g), AC+$K$-Means for $L = 3$ (h), $L = 6$ (i), AR+$K$-Means for $P = 3$ (j), $P = 6$ (k), and $K$-AR for $P = 2$ (l), $P = 3$ (m), $P = 4$ (n), $P = 6$ (o) and $P = 10$ (p).

limitations of the five algorithms to identify defected and cracked regions. In figures 4.24 (h), (i), (j) and (k), AC+$K$-means and AR+$K$-means produce better segmented regions. The main defected regions are better identified for two different values of $L$ and $P$, while some cracks and details are not well extracted inside the cork. Figures



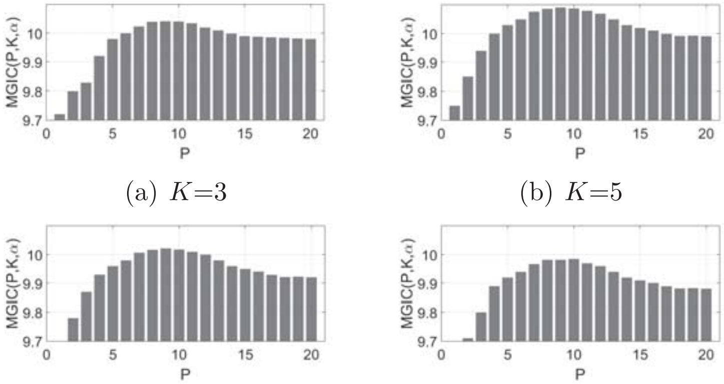
(a) $K$=2

(b) $K$=4

(c) $K$=6

(d) $K$=8

Figure 4.26 – Modified generalized information criterion (MGIC) computed after the chemical THz image segmentation for different values of parameters $P$ and $K$.



(a) $K$=3

(b) $K$=5

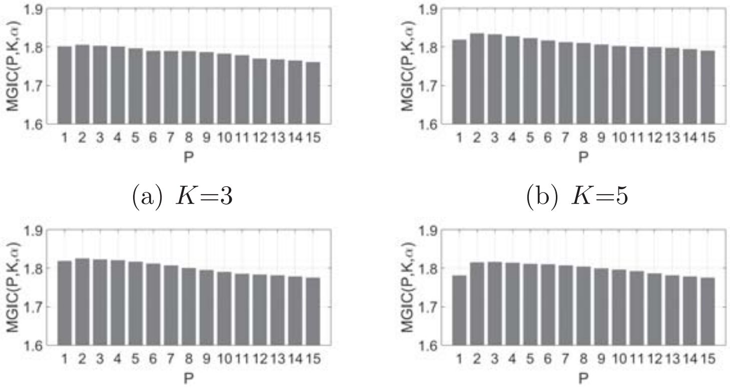
(c) $K$=7

(d) $K$=9

Figure 4.27 – Modified generalized information criterion (MGIC) computed after the cork THz image segmentation for different values of parameters $P$ and $K$.

4.24 from (l) to (p) display the obtained regions of $K$-AR for $P$ equal to 1, 4, 8, 10, and 12. The cork grains, the voids, the defects and the cracks are well segmented for different values of $P$, mainly for $P > 4$ and $P < 12$.

Figure 4.25 shows the segmentation outputs of $K$-means, KHM, GMM, W-$K$-means, SS-$K$-means, AC+$K$-means, AR+$K$-means and $K$-AR clustering algorithms on the moth Terahertz images. $K$-means, KHM, GMM, W-$K$-means and SS-$K$-means with $c < 2$ produce wrongly segmented regions in figures 4.25 (a), (b), (c), (d), (e)



(a) $K$=3

(b) $K$=5

(c) $K$=7

(d) $K$=9

Figure 4.28 – Modified generalized information criterion (MGIC) computed after the moth THz image segmentation for different values of parameters $P$ and $K$.

and (f). SS-$K$-means with $c \geq 2$ produce better segmented regions in figure 4.25 (g). However, moth body structure are not yet identified. The obtained regions clearly illustrate the limitations of the five techniques to provide good inner structure of the body and the wings. In figures 4.25 (h) and (i), AC+$K$-means produce better segmented regions. The moth wings are well identified with AC+$K$-means for $L = 3$ and $L = 6$, while the moth body is not well segmented. In figures 4.25 (j) and (k), AR+$K$-means produces noisy segmented regions. The moth wings and the body are not well segmented. Figures 4.25 from (l) to (p) display the obtained regions of $K$-AR for $P$ equal to 2, 3, 4, 6, and 10. The structure of the moth wings is well segmented for different values of $P$ and the moth wings and the body are preserved.

Figures 4.26, 4.27 and 4.28 show plots of the MGIC criterion for divers orders of $K$-AR approach on the three THz images. The parameter $\alpha$ is fixed to $2(log(log(T+N)))$. In figure 4.26, MGIC values are high for $P > 8$ and the highest one corresponds to the $P = 13$. As previously shown in figures 4.20, 4.24 and 4.25, accuracy, precision and recall measures are high for $P > 4$ and the highest is for $P = 13$ which corresponds to the value of parameter $P$ selected by MGIC criterion. As already seen in section 2, the computational complexity of $K$-AR algorithm is quadratic for parameter $P$. So, high values of $P$ are not preferable and a suitable interval between 5 and 12 can be also interesting. With a such value of $\alpha$, the MGIC criterion allows to select the AR order having the highest clustering performance with moderate and reasonable AR order. In addition, the selection of the cluster number corresponds to the maximal value of MGIC with respect to $K$. So, $K = 4$ corresponds to the correct number of chemical components. Figure 4.27 show that the maximal values of MGIC correspond to $K = 5$. Also, the values of $P$ between 5 to 12 correspond to the highest values of MGIC and in the same time to the best segmentation of the cork THz image (see figures 4.24 from (l) to (p)). The estimated values of parameters $P$ and $K$ show the best compromise between high performances and simplicity (low number of parameters). The maximal value of MGIC corresponds to $P = 9$ which corresponds to the best detection of the defects and the cracks inside the cork. Figure 4.28 shows that the maximal values of MGIC correspond to $K = 5$. Also, the values of $P$ between 2 to 6 correspond to the highest values of MGIC and in the same time to the best segmentation of the moth THz image (see figures 4.25 from (l) to (p)). The estimated values of parameters $P$ and $K$ show the best compromise between high clustering performances and low number of parameters. The maximal value of MGIC corresponds to $P = 2$ which corresponds to the best moth wings and the body identification.

# 6 Conclusion

In this paper, we have proposed a novel clustering approach, called $K$-AR model and suitable to THz images based time series. The $K$-AR approach consists to regroup a set of THz time series into clusters represented by their prediction prototypes. The $K$-AR assumes that the time series depicting the pixels were generated by AR models and consists to recover the original $K$ autoregressive models describing each cluster of time series. The corresponding pixels are then assigned to the clusters having the best AR model fitting. The order and the number of clusters of $K$-AR model are automatically estimated using a modified information criterion. Our approach is tested on various artificial and real THz images. Experimental results show that $K$-AR approach produces more accurate segmentation than other clustering techniques such as $K$-means, GMM, KHM, W-$K$-means, SS-$K$-means, AC+$K$-means and AR+$K$-means.

Our approach is shown so attractive to achieve the best clustering performances.

Note that the sensitivity to initial starting conditions decreases the clustering accuracy. Furthermore, our clustering approach use hard decision to compute the membership of pixels and deal all the features with equal importance. Soft decision rule and feature weighting techniques can improve the accuracy of the analysis. These problems haven't been addressed in this paper and require further studies.

In further work, we will deal with the initialization of the weights of AR models and the weights of the features. Also, we will extend our approach to classify non-stationary time series which cannot be fitted by AR linear models; also, partition the set of time series in the frequency domain. Similar steps to the work of Maharaj and D'Urso [79] about the time series clustering in the frequency domain, one can improve the Terahertz image analysis.

## Acknowledgments

The authors would like to thank the editor and the three referees for their useful comments which helped to improve the quality of this manuscript. The authors thank also Professor John F. Federici of New Jersey Institute of Technology USA and Professor Thomas Tongue of Zomega Terahertz Corporation for the Terahertz measurements.

# Chapitre 5

# Mélange fini de modèles autorégressifs et ses applications pour la classification des séries chronologiques

Dans ce chapitre, nous proposons une généralisation de l'approche présentée dans le chapitre précèdent. Au lieu de considérer un problème de moindres carrés, nous proposons une approche de classification probabiliste basée sur le mélange de modèles autorégressifs. L'approche proposée consiste à récupérer les modèles autorégressifs originaux décrivant chaque distribution de séries chronologiques. L'estimation par la méthode de maximum de vraisemblance est utilisée pour apprendre les paramètres de l'approche proposée. L'ordre du modèle autorégressif et le nombre de composants du mélange sont automatiquement estimés en utilisant un critère de sélection du modèle.

Dans ce chapitre, nous présentons un article intitulé **Finite mixture of autoregressive models and its applications in time series clustering** soumis dans le journal international de Elsevier **Engineering Applications of Artificial Intelligence**. Le problème a été posé par le professeur Djemel Ziou. J'ai réalisé, validé et rédigé ce travail sous sa supervision.

# Finite mixture of autoregressive models and its applications in time series clustering

Mohamed Walid Ayech
Département d'informatique, Université de Sherbrooke,
Sherbrooke, Québec, Canada J1K 2R1
walid.ayech@usherbrooke.ca

Djemel Ziou
Département d'informatique, Université de Sherbrooke,
Sherbrooke, Québec, Canada J1K 2R1
djemel.ziou@usherbrooke.ca

**Keywords**: Clustering, autoregressive model, maximum likelihood estimation, model selection criterion, Terahertz imaging.

## Abstract

We propose in this paper a new clustering approach based on autoregressive techniques. The proposed approach is called mixture of autoregressive (MoAR) models and assumes that the time series were generated by autoregressive models. MoAR approach consists to recover the original autoregressive models describing each cluster of time series. The parameters of MoAR model are automatically estimated by using a model selection criterion. Our approach is tested on various synthetic datasets, robotic datasets, transit events and Terahertz images. Experimental results show that MoAR approach is more efficient than other clustering approaches from the literature.

# 1 Introduction

In the last years, time series clustering has attracted a lot of interest in the scientific community [2, 46, 73]. It is essential to find solutions for real problems deriving from several domains of applications such as pharmaceutics, finance, environmetrics, genetics and bioinformatics [2]. For instance, in finance, time series clustering has been used to group companies listed in the market exchange by examining their time series of returns [47]. In pharmaceutics, it is used to classify temporal responses of drugs provided from patients after taking it [48]. In genetics, time series clustering has been used to group genes taking into account profiles of temporal expression from cDNA microarrays experiments [84]. In astronomy, it has been used to classify time series of star brightness in huge data and to estimate prototype stars [47]. In biomedicine, the study of EEG biological signals requires to discriminate between signals caused by sick or healthy people [87]. In image processing, it has been used for texture analysis [62]. Several other applications of time series clustering are detailed in [2].

In general, a time series is a sequence of observations related in chronological order. Some authors consider it as a vector represented by high number of bands or features. They found useful to extract lower dimensional features before classifying the time series. Some useful features are extracted by using autocorrelation functions [48], spectral transformations [79], wavelets coefficients [44, 74] and reconstructed phase spaces [92]. Other authors assume that the values associated with each time series are sampled from an autoregressive (AR) models [48, 14] or autoregressive moving average (ARMA) models [18, 109]. The clustering of the set of time series is realized in the extracted parameter spaces. The advantages of these methods are that the clustering is realized in low dimensional feature space and the correlation property is incorporated into the process of time series clustering. However, there are several challenges in analyzing time series. The choice of the distribution of residues (e.g. Gaussian distribution [27], mixture of Gaussians [91] or Beta distribution [32]) must be well identified. The time series must be stationarity before the analysis process. The choice of the model (e.g. AR, MA, ARMA or ARIMA) and the order of the model must be suitably selected before estimating the parameters of the model. Moreover, the methods of parameter estimation are numerous (e.g. least squares estimator, Yule-

Walker estimator, maximum likelihood estimator and Bayesian estimator). What is the best method suitable for time series fitting? The model parameters are estimated from each time series, independently, and the interpretation of these parameters and the space which constitute is not easy [1]. For example, what is the mean of the first coefficients of the AR model?

Several clustering methods are proposed in the literature. Among them, $K$-means [48], $K$-medoids [18], $K$-harmonic-means [14] and agglomerative hierarchical [78]. Other works use Gaussian mixture models (GMM) to discriminate between the time series [74, 92]. In GMM, each component is modeled by the multivariate Normal distribution and it is characterized by its mean vector and its covariance matrix. The covariance matrix allows to determine the geometric features (volume, shape and orientation) of each component of time series. However, the time series are characterized by a dynamic behaviour in their evolution over time [48]. This dynamic behaviour can be defined by the correlation property into the time series and it is not incorporated into the process of standard GMM clustering. Instead, we believe to model each cluster of time series by a single univariate AR model. The AR parameters must be estimated by using both the correlation property between time series variables and the structure of the time series clusters. We look forward an iterative algorithm which alternate between autoregressive modeling and time series clustering until reaching a high time series fitting. In this paper, we introduce a novel clustering approach called mixture of autoregressive (MoAR) models which consists to classify the set of time series into a fixed number of clusters. We assume that each time series is stationary and generated from an univariate autoregressive (AR) model where the order is unknown. A model-based method is proposed for recovering the original autoregressive models describing each cluster of time series. These time series are then assigned to the clusters having the best AR model fitting. The parameters of the MoAR model are automatically estimated by using a modified generalized information criterion. The proposed approach is tested to discriminate transient events for a safe monitoring process, detect the surface nature of a mobile robot, and segment various Terahertz images. There are three main differences between our work and the state of art. First, our work constitutes a new clustering method based on autoregressive techniques which is more adapted to time series and not excessively affected by outliers. Second, the time series components take into account the dynamic behaviour of the time series. Third, our approach is completely unsupervised, the parameters of MoAR model are automatically estimated by using a modified generalized information criterion.

The rest of the paper is organized as follows: in section 2 we introduce the AR for univariate signal. Section 3 presents an original approach, called MoAR models, for time series clustering. The selection of the MoAR model is proposed in section 4. The results are illustrated and discussed in section 5.

# 2 Autoregressive modeling

Let us consider a discrete-time random process $X = \{X_1, \dots, X_T\}$. We assume that $X$ is a stationary and ergodic process and a data $x = (x_1, \dots, x_T)^{'}$ is generated from this process. Moreover, a random variable $X_t$ can be defined by a linear combination of $X_{t-1}, \dots, X_{t-P}$ plus a random variable representing the bias often assumed a centered Normal. In this case, the process is autoregressive (AR) known as time series of order $P$. The complexity of time series is equal to $P+1$, where one accounts for the variance of the bias. More formally, AR model predicts the next value $x_t$ in the time series as following

$$x_t = \phi_t^{'} w + \varepsilon_t, \tag{5.1}$$

where $\phi_t = (x_{t-1}, x_{t-2}, \dots, x_{t-P})^{'}$ represents a vector of the $P$ previous realizations of $x_t$, $w = (w_1, \dots, w_P)^{'}$ is a column vector of AR weights and $\varepsilon_t$ is Gaussian noise with zero mean and $\sigma^2$ variance. Given a set of parameters $\theta = \{w, \sigma^2\}$, the likelihood of the sequence of residuals is given by

$$p(\varepsilon|\theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{T-P}{2}} exp\left(-\frac{1}{2\sigma^2}\sum_{t=P+1}^{T}\left(x_t - \phi_t^{'} w\right)^2\right) \cdot \tag{5.2}$$

The AR weights $w$ and the noise variance $\sigma^2$ will be determined by fitting the model to the training time series data. This problem can be solved by a large variety of techniques such as the least squares method, the Yule-Walker method, the method of moments and the maximum likelihood method. The parameters $w$ and $\sigma^2$ are estimated in our work by using a maximum likelihood (ML) solution which consists to maximize the logarithm of the likelihood function in equation (5.2). We have

$$\log p(\varepsilon|\theta) = -\frac{T-P}{2}\log\sigma^2 - \frac{T-P}{2}\log(2\pi) - \frac{E(w)}{\sigma^2}\cdot \tag{5.3}$$

Consider first the maximization with respect to $w$, the gradient of the log-likelihood function written in equation (5.3) takes the form

$$\frac{\partial}{\partial w}\log p(\varepsilon|\theta) = \frac{1}{\sigma^2}\sum_{t=P+1}^{T}(x_t - \phi_t^{'}w)\phi_t = 0\cdot \tag{5.4}$$

We can deduce the maximum likelihood solution for the AR coefficients as follows

$$w = (\phi^{'}\phi)^{-1}\phi^{'}x, \tag{5.5}$$

where $\phi$ is a $(T-P)\times P$ matrix containing the row vectors $\phi_t$. By using the estimated autoregressive weights $w$, the maximum likelihood noise variance $\sigma^2$ can be estimated from

$$\sigma^2 = \frac{1}{T-P}\sum_{t=P+1}^{T}(x_t - \phi_t^{'}w)^2\cdot \tag{5.6}$$

In the next section, we will formulate the time series clustering by using a mixture of AR models.

# 3 The proposed MoAR models

The error $\varepsilon_t$ is often considered as a Gaussian random variable [27]. Some authors considered it as Beta random variable [32]. These hypotheses are especially justified by the simplicity of the design of a parameter estimator. Motivated by the simplicity of the estimators, one can also think of considering the error as a mixture of pdfs. This idea has been implemented in the state of the art in [91]. Indeed, the mixture of Gaussians was used to classify time series [74, 92, 2]. It was generally described by the raw feature space and used without considering the correlation property between the time series variables [41, 81]. We believe that autoregressive (AR) modeling leads to model each cluster of time series. In this section, we propose a novel clustering approach called mixture of autoregressive (MoAR) models, which consists to classify the set of time series into a fixed number of clusters. We assume that each time series is stationary and generated from an AR model, and the set of time series are arisen from a mixture of components in different proportions. The model order and the number of components are initially specified. MoAR approach consists to classify the AR models and recovers the original $K$ autoregressive models describing each cluster of time series.

Let us consider $N$ discrete-time random process $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_N\}$. Since each $\mathcal{X}_n = \{X_{n1}, \ldots, X_{nT}\}$ is an ergodic process, its realizations correspond to the $n^{th}$ time series $x_n = (x_{n1}, \ldots, x_{nT})^{'}$ where $T$ is the number of realizations. Each cluster of time series is depicted by an AR model characterized by its weights $w_k = (w_{k1}, \ldots, w_{kP})^{'}$, where $P$ is the order of the model. $\phi_{nt} = (x_{n(t-1)}, x_{n(t-2)}, \ldots, x_{n(t-P)})^{'}$ is a vector of the $P$ previous realizations of $x_{nt}$. As described in Section 2, the current value $x_{nt}$ of each time series $x_n$ can be written as a weighted linear sum of its previous values and given by $\phi^{'}_{nt} w_k + \varepsilon_{nt}$, where $\varepsilon_{nt}$ is additive Gaussian noise with zero mean and $\sigma_k^2$ variance.

The process of clustering consists to assign each time series $x_n$ to the clusters having the best AR model fitting. Let us define the multinomial random vector $\mathcal{Z}_n = \{Z_{n1}, \ldots, Z_{nK}\}$ associated with $x_n$, where its realizations $z_n = (z_{n1}, \ldots, z_{nK})^{'}$ indicate the index of the class assigned to it. Each $z_{nk} \in \{0, 1\}$, $\sum_{k=1}^{K} z_{nk} = 1$, and $z_{nk} = 1$ if the time series $x_n$ belong to cluster $k$ and 0, otherwise. We defined the pdf $p(x_n, z_n|\theta, \pi) = p(z_n|\pi)p(x_n|z_n, \theta)$, where $p(z_n|\pi) = \pi_k$ and $p(x_n|z_n, \theta) = \mathcal{N}(x_n|\phi_n w_k, \sigma_k)$. $\mathcal{N}(x_n|\phi_n w_k, \sigma_k)$ represents the Normal distribution of the component $k$, $\phi_n$ is a $(T-P) \times P$ matrix containing the row vectors $\phi_{nt}$, $\theta$ represents the set of parameters $\{w_1, \cdots, w_K, \sigma_1, \cdots, \sigma_K\}$ and $\pi = \{\pi_1, \cdots, \pi_K\}$ represents the mixing coefficients. For all time series, the conditional distribution of the random vector $\mathcal{Z} = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_N\}$, given the mixing coefficients $\pi$, can be then specified in the following form

$$p(\mathcal{Z}|\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{5.7}$$

and the conditional distribution of $\mathcal{X}$ given the random vector $\mathcal{Z}$ and the set of parameters $\theta$ is

$$p(\mathcal{X}|\mathcal{Z},\theta) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(x_n|z_n,\theta_k)^{z_{nk}}. \tag{5.8}$$

The joint distribution $p(\mathcal{X},\mathcal{Z}|\theta,\pi)$ is then given by

$$p(\mathcal{X},\mathcal{Z}|\theta,\pi) = \prod_{n=1}^{N}\prod_{k=1}^{K} (\pi_k p(x_n|\theta_k))^{z_{nk}}. \tag{5.9}$$

The associated mixture of pdfs is given by

$$L(\Theta) = \prod_{n=1}^{N}\left(\sum_{k=1}^{K}\pi_k \mathcal{N}(x_n|\phi_n w_k,\sigma_k^2)\right). \tag{5.10}$$

The proposed model is specified by the set of parameters $\Theta = \{\theta,\pi\}$, $K$ and $P$. The best model must be selected in which the goal is to achieve the best tradeoff between model fitting and model complexity. In the following section, we present how different parameters are estimated via a model selection criterion.

# 4 Model selection

In this section, we propose to reformulate the time series clustering as a model selection problem. The selection of the model is generally realized by using information criteria in which the goal is to achieve the best tradeoff between model fitting and model complexity [7, 98, 49, 30]. A comparison between some model selection criteria is detailed in [25]. Generalized information criterion (GIC) [30] is an objective measure largely used for model selection. In our work, we propose a modified generalized information criterion (MGIC) which is well adapted to a set of time series. The proposed criterion consists to estimate the best MoAR model. It is defined as follows:

$$MGIC(P,K,\alpha) = 2\ln L(\Theta) - \alpha(P+2)K \tag{5.11}$$

where $\Theta = \{\pi, w, \sigma\}$ represents the set of parameters of MoAR, $(P+2)K$ represents the number of parameters to be estimated, $\alpha$ is a positive real penalty factor and

$\ln L(\Theta)$ represents the log-likelihood function which is given as following

$$\ln L(\Theta) = \sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n | \phi_n w_k, \sigma_k^2) \right) \cdot \tag{5.12}$$

Maximizing MGIC criterion return to maximizing the log-likelihood of the model and minimizing the number of the used parameters. To select a model order and a mixture component number, MGIC($P$,$K$,$\alpha$) criterion is determined for all $P$ between one and a maximum candidate order $P_{max}$ and for all $K$ between two and a maximum candidate number $K_{max}$. $P_{max}$ and $K_{max}$ are fixed beforehand. The order and the component number with the maximal value of the criterion are selected. To select the parameters $\Theta = \{\pi, w, \sigma\}$, the log-likelihood function given by equation (5.12) is maximized by using the maximum likelihood (ML) estimate associated with a sample of time series [94, 41, 81, 10]. The expectation maximization (EM) is an iterative algorithm which has been suggested as alternative way to find maximum likelihood solutions for models having latent variables [41]. EM algorithm for a mixture of AR models is applied on the given data set $\mathcal{X}$. In the E-step, the posterior distribution of $\mathcal{Z}$ is called $\mathcal{R} = \{r_1, \ldots, r_N\}$, where each $r_n = (r_{n1}, \ldots, r_{nK})'$ and each $r_{nk}$ represents the responsibilities or the assignments of the $n^{th}$ time series to the $k^{th}$ component. This posterior distribution is computed by using the current values of the parameters $\Theta$ and it is given by

$$r_{nk} = \frac{\pi_k \mathcal{N}(x_n | \phi_n' w_k, \sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n | \phi_n' w_j, \sigma_j)} \cdot \tag{5.13}$$

In the M-step, the log-likelihood is maximized with respect to the parameters $\Theta$. Taking the corresponding partial derivatives $\partial L(\Theta)/\partial w_k$ equal to zero, we find the following relationship as a function of $w_k$

$$\sum_{n=1}^{N} \left( \left( \phi_n' x_n - w_k \phi_n' \phi_n \right) \times \frac{\pi_k \mathcal{N}(x_n | \phi_n w_k, \sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n | \phi_n w_j, \sigma_j)} \right) \cdot \tag{5.14}$$

By using equation (5.13), we can then derive the AR weights as following

$$w_k = \left( \sum_{n=1}^{N} r_{nk} \phi_n' \phi_n \right)^{-1} \left( \sum_{n=1}^{N} r_{nk} \phi_n' x_n \right), \tag{5.15}$$

where $\phi_n$ is a $(T-P) \times P$ matrix containing the row vectors $\phi_{nt}$. Note that the estimated weights is a multiplication of two terms. Equation (5.15) shows that AR weights are estimated by using the structure of time series classes represented by $r_{nk}$ and the empirical correlated variables represented by both terms $\phi_n^{'}\phi_n$ and $\phi_n^{'}x_n$. The log-likelihood is also maximized with respect to $\pi$. Taking the corresponding partial derivatives equal to zero and using the Lagrange multipliers for the constraint $\sum_{k=1}^{K}\pi_k = 1$, we can derive the mixing coefficients as following

$$\pi_k = \frac{N_k}{N}, \tag{5.16}$$

where $N_k = \sum_{n=1}^{N} r_{nk}$ and represents the number of time series in the $k^{th}$ component. In similar way, the noise variance $\sigma_k^2$ can be updated by using the following equation

$$\sigma_k^2 = \frac{1}{N_k(T-P)} \sum_{n=1}^{N} \sum_{t=P+1}^{T} r_{nk}(x_{nt} - \phi_{nt}^{'} w_k)^2. \tag{5.17}$$

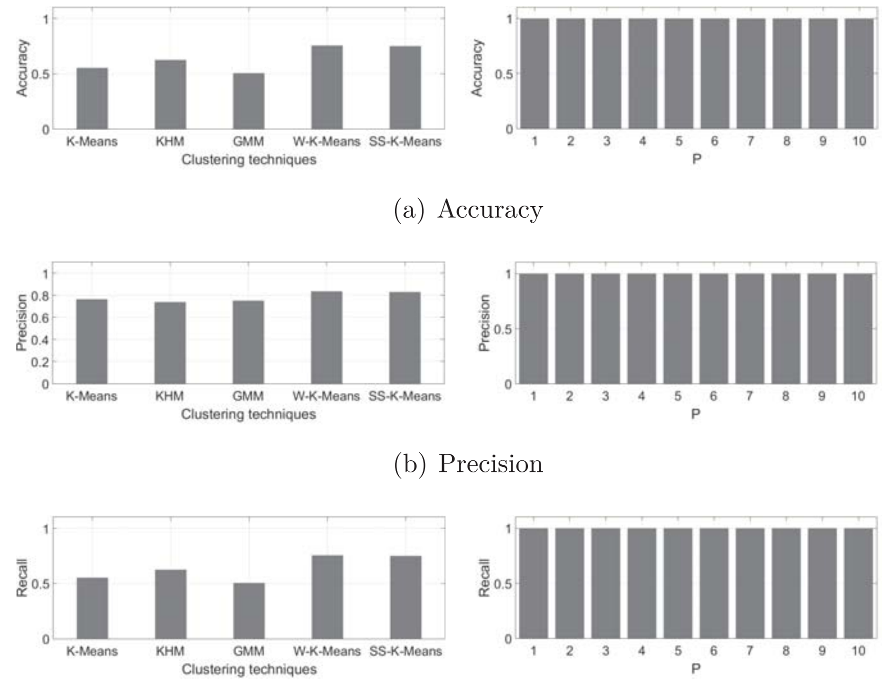
The learning process is then done by iterating between E and M steps, i.e. updating the responsibilities $r_{nk}$ of the time series, the AR weights $w_k$, the mixing coefficients $\pi_k$ and the noise variance $\sigma_k^2$ until the value of likelihood $L(\Theta)$ of the data will be minimal.

# 5 Experimental results

In this section, various synthetic and real datasets are used to investigate the performance of the clustering approaches. We consider four case studies in different domains. We experimentally and empirically show that MoAR outperforms other approaches such as Gaussian mixture model (GMM) [41], standard $K$-means [77], $K$-harmonic-means (KHM) [111], W-$K$-means [64] and SS-$K$-means [13].

## 5.1 Experiments on synthetic datasets

In this experiment, we used three artificial datasets with various cluster number to verify the performances of the clustering algorithms. The three datasets $D_P$ were



(a) Accuracy

(b) Precision

(c) Recall

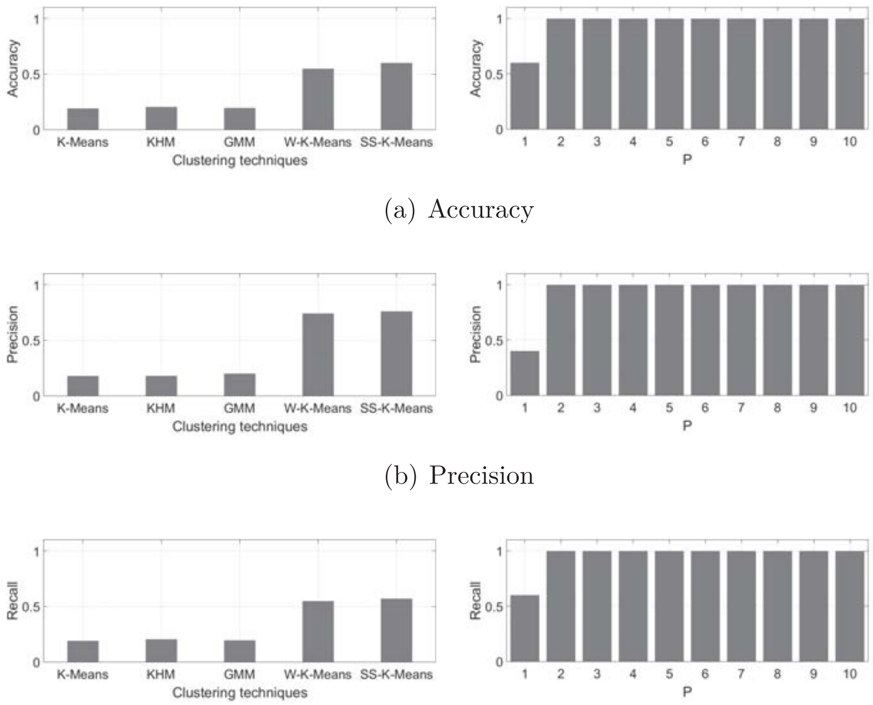
Figure 5.1 – Clustering performances on dataset $D_2$ for $K$-means, KHM, GMM, W-$K$-means and SS-$K$-means (in the left) and MoAR for $P$ from 1 to 10 (in the right). The parameter $K$ is fixed to 2.

constituted by $K$ clusters of time series originated from univariate AR($P$) process using the following equation

$$x_{nt} = \phi_{nt}^{'} w_n + e_{nt} \tag{5.18}$$

where $x_{nt}$ is the $t^{th}$ observation of the $n^{th}$ time series, $w_n$ is a column vector of $P$ AR weights for the $n^{th}$ time series, $\phi_{nt} = (x_{n(t-1)}, x_{n(t-2)}, \ldots, x_{n(t-P)})'$ are the $P$ previous time series observations of $x_{nt}$ and $e_{nt}$ is additive Gaussian noise with zero mean and $\sigma_n^2$ variance. For the different time series, the variances of the Gaussian noise are randomly chosen between 0.001 and 0.1.

In these experiments, data set $D_2$ was constituted by 400 stationary time series distributed into two clusters originated from univariate AR(2) process. Each time series is characterized by 500 features. The AR weights $w_n$ are equal to $(1.8, -.9)'$



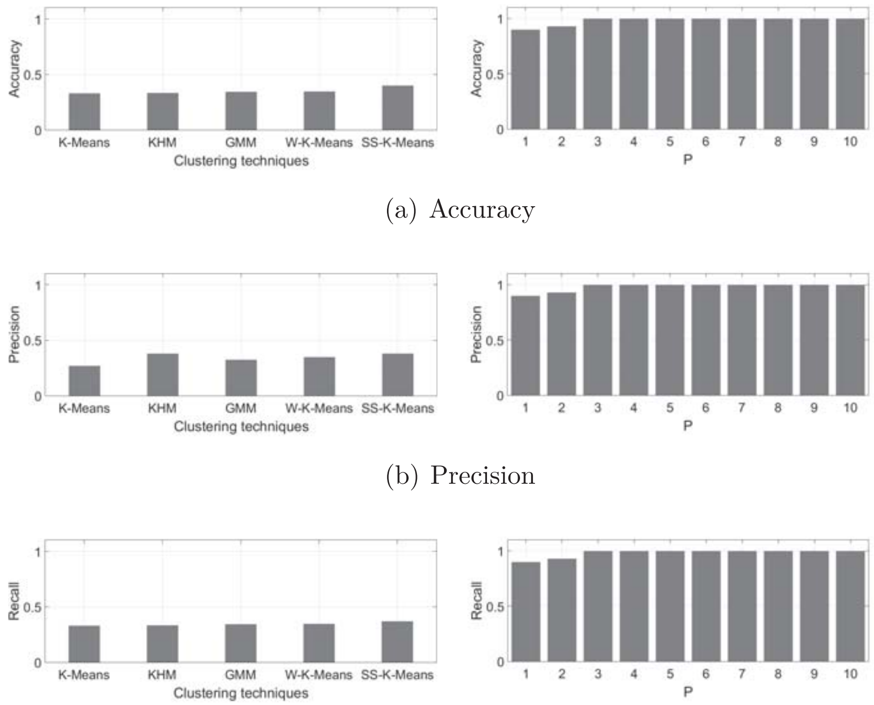
(a) Accuracy

(b) Precision

(c) Recall

Figure 5.2 – Clustering performances on dataset $D_3$ for $K$-means, KHM, GMM, W-$K$-means and SS-$K$-means (in the left) and MoAR for $P$ from 1 to 10 (in the right). The parameter $K$ is fixed to 5.

and $(.5, -.3)'$ respectively for time series of cluster 1 and cluster 2. Data set $D_3$ was constituted by 300 stationary time series distributed into five clusters originated from univariate AR(3) process. Each time series is characterized by 1000 features. The AR weights $w_n$ are equal to $(-.6, -.1, -.1)'$, $(.005, 0, -.001)'$, $(.1, -.5, .5)'$, $(.5, .2, .2)'$ and $(1, -.6, .2)'$ respectively for cluster 1, cluster 2, cluster 3, cluster 4 and cluster 5. Data set $D_5$ was constituted by 500 stationary time series distributed into three clusters originated from univariate AR(5) process. Each time series is characterized by 500 features. The AR weights $w_n$ are equal to $(-.3, -.1, -.1, .1, -.1)'$, $(.005, 0, -.001, -.005, .005)'$ and $(.3, .1, .2, -.2, -.2)'$ respectively for cluster 1, cluster 2 and cluster 3. The different clustering techniques were statistically compared in terms of clustering performance. We used accuracy, precision and recall to evaluate the results [93]. The accuracy is the proportion of time series correctly classified. The



(a) Accuracy

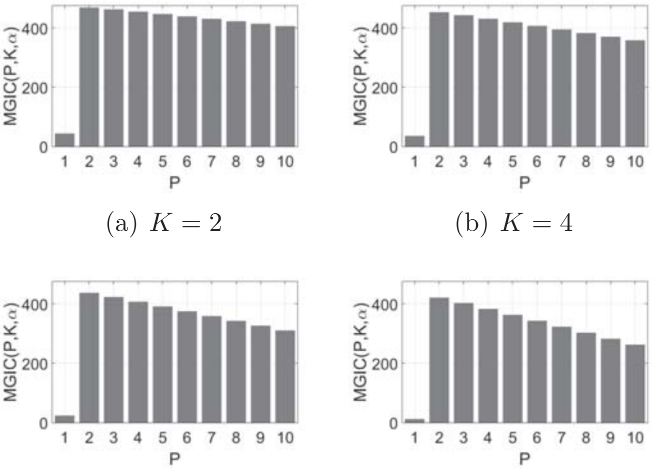
(b) Precision

(c) Recall

Figure 5.3 – Clustering performances on dataset $D_5$ for $K$-means, KHM, GMM, W-$K$-means and SS-$K$-means (in the left) and MoAR for $P$ from 1 to 10 (in the right). The parameter $K$ is fixed to 3.

precision is computed as the fraction of classified time series which belong to the relevant class. The recall is computed as the fraction of the relevant time series which are correctly classified.

This section consists to evaluate the performances of the clustering techniques ($K$-means, KHM, GMM, W-$K$-means and SS-$K$-means) and the MoAR for different values of the parameter $P$. SS-$K$-means was used with parameters $a = 2$, $b = 2$ and $c = 1$. Figures 5.1, 5.2 and 5.3 comprise the evaluation measures for the three artificial datasets. For data set $D_2$, the clustering accuracy is less than 0.6 for $K$-means, GMM and KHM and around 0.75 for W-$K$-means and SS-$K$-means. While, it is equal to 1.0 for MoAR clustering for $P$ from 1 to 10. The precision measure is around 0.75 for $K$-means, GMM and KHM, around 0.8 for W-$K$-means and SS-$K$-means, and equal to 1.0 for MoAR for different values of $P$. Also, recall measure does not surpass 0.65



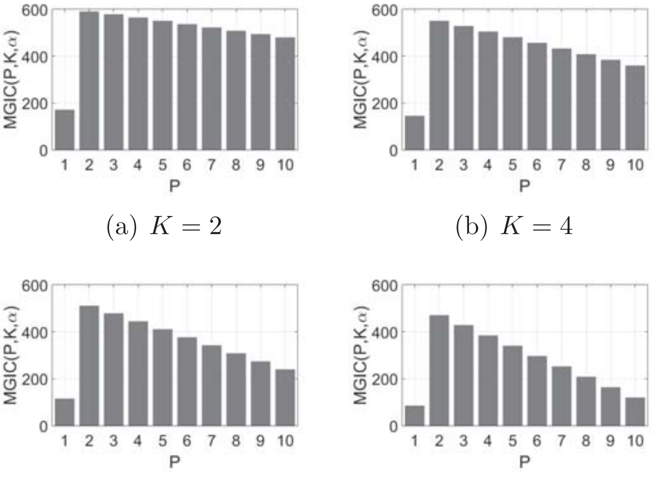
(a) $K = 2$

(b) $K = 4$

(c) $K = 6$

(d) $K = 8$

Figure 5.4 – MGIC criterion for dataset $D_2$. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ and $P$. The parameter $\alpha$ is fixed to 2.



(a) $K = 2$

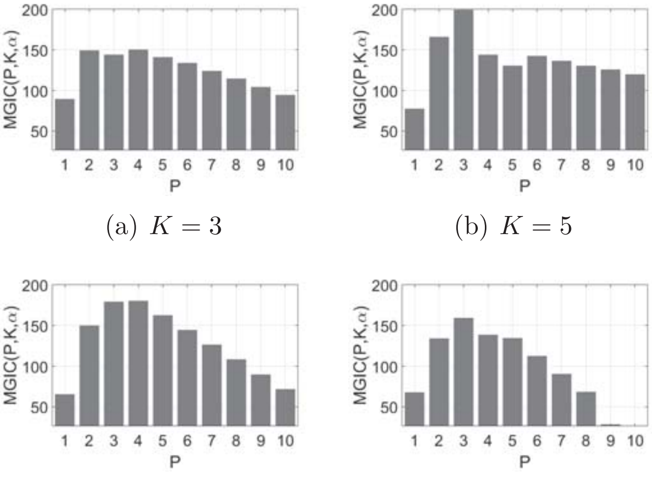
(b) $K = 4$

(c) $K = 6$

(d) $K = 8$

Figure 5.5 – MGIC criterion for dataset $D_2$. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ and $P$. The parameter $\alpha$ is fixed to $2 \ln \ln(TN)$.

for $K$-means, GMM and KHM, around 0.75 for W-$K$-means and SS-$K$-means and equal 1.0 for MoAR clustering for different values of $P$. In figures 5.2 (a), (b) and (c), clustering accuracy, precision and recall of artificial dataset $D_3$ are around 0.2



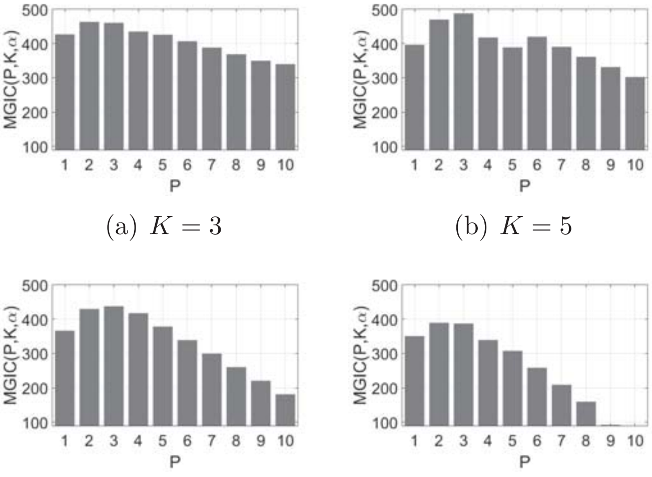
(a) $K = 3$

(b) $K = 5$

(c) $K = 7$

(d) $K = 9$

Figure 5.6 – MGIC criterion for dataset $D_3$. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ and $P$. The parameter $\alpha$ is fixed to 2.



(a) $K = 3$

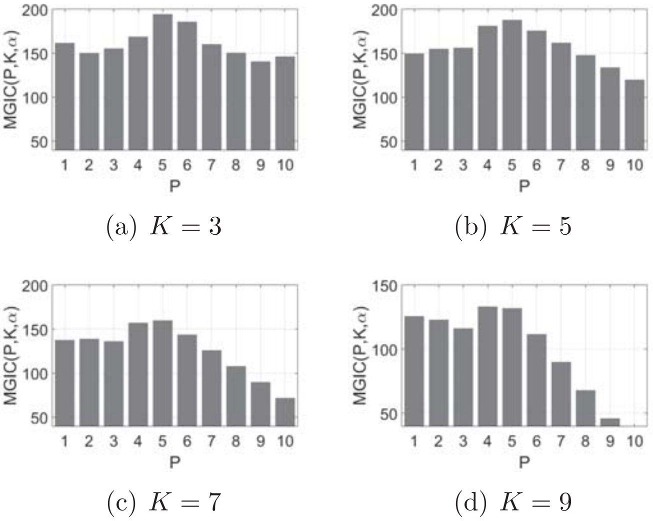
(b) $K = 5$

(c) $K = 7$

(d) $K = 9$

Figure 5.7 – MGIC criterion for dataset $D_3$. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ and $P$. The parameter $\alpha$ is fixed to $2 \ln \ln(TN)$.

for $K$-means, GMM and KHM and does not surpass 0.75 for W-$K$-means and SS-$K$-means. While, they are between 0.4 and 0.6 for MoAR with $P = 1$ and 1.0 for MoAR with $P > 1$. In figures 5.3 (a), (b) and (c), clustering accuracy, precision and recall



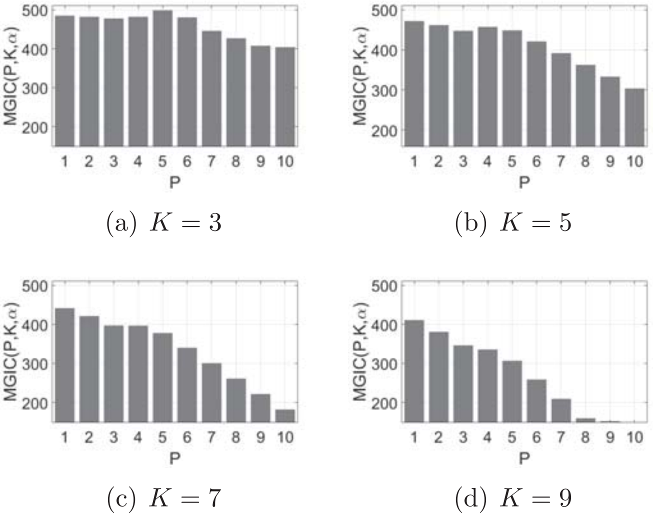
(a) $K = 3$

(b) $K = 5$

(c) $K = 7$

(d) $K = 9$

Figure 5.8 – MGIC criterion for dataset $D_5$. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ and $P$. The parameter $\alpha$ is fixed to 2.



(a) $K = 3$

(b) $K = 5$

(c) $K = 7$

(d) $K = 9$

Figure 5.9 – MGIC criterion for dataset $D_5$. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ and $P$. The parameter $\alpha$ is fixed to $2 \ln \ln(TN)$.

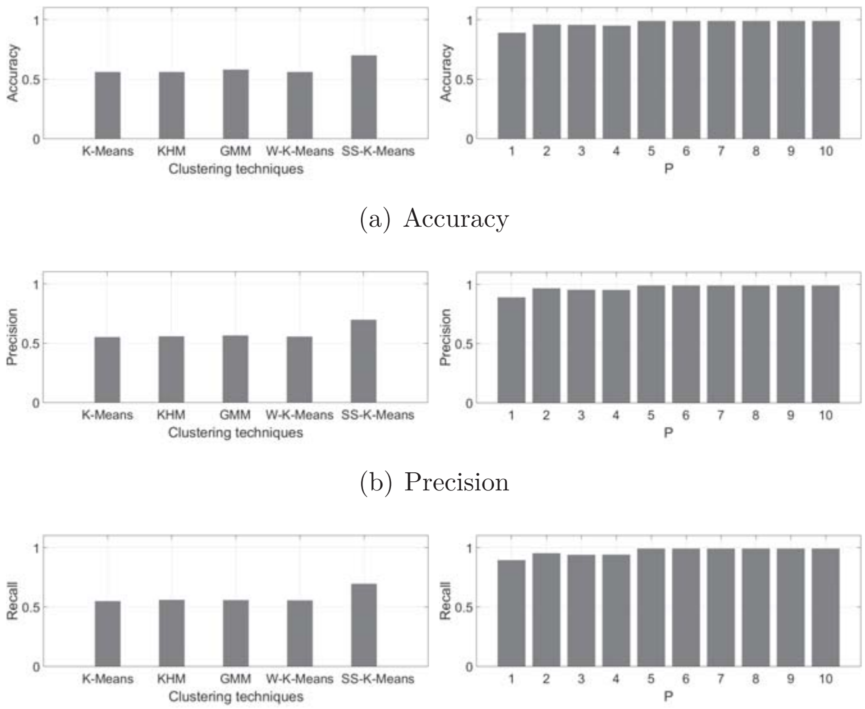
of artificial dataset $D_5$ are between 0.25 and 0.35 for $K$-means, GMM, KHM, W-$K$-means and SS-$K$-means. These measures are around 0.85 for MoAR with $P = 1$, 0.9 for MoAR with $P = 2$ and become 1.0 for MoAR when $P > 2$. The obtained statistics show the high performances of MoAR compared to the other approaches.

Mixture component number $K$ and AR model order $P$ are selected by using MGIC($P$,$K$,$\alpha$) criterion. In this paper, we study the parameter selection of our clustering approach for two different values of $\alpha$ ($\alpha = 2$ and $\alpha = 2\ln\ln(TN)$). MGIC with both values of $\alpha$ can be considered as modified versions of Akaike information criterion [7] and Hannan-Quinn information criterion [49], respectively. Figures from 5.4 to 5.9 show the corresponding MGIC measures of MoAR($P$) as functions of $P$ for four different values of $K$. These figures show that MGIC values give a correct AR model order estimation for different cluster number. The highest MGIC measure corresponds to $K$ equal to 2, 5 and 3 and $P$ equal to 2, 3 and 5, respectively for datasets $D_2$, $D_3$ and $D_5$. These figures show that $P$ and $K$ are well estimated by MGIC criterion and confirm the results found previously in figures 5.1, 5.2 and 5.3.

## 5.2 Recognition of transient events

Several industrial processes are carried out in long periods of steady-state running. These states are usually interspersed with shorter periods with a nature more dynamic corresponding to abnormal events, called transient events. TRACE dataset is a sample of transient events classification Benchmark [36]. It is obtained from EDF (Electricité de France, of the PWR 900 MW nuclear power plant), designed to simulate different classes of transit events (transitions to different operation states, major disturbances, actuator failures and instrumentation failures) in a nuclear power plant, produced in the form of time series data and reported by Davide Roverso [97]. Recognition of transient events is a challenge for the safe and economical operation of the monitored process.

TRACE dataset contains 200 instances distributed into 4 classes, 50 for each class. All instances are linearly interpolated and normalized to have the same length of 275 data points. TRACE dataset is tested by the different clustering techniques which are statistically compared in terms of clustering accuracy, precision and recall. Figure 5.10 shows the quantitative performances of the obtained clusters on the TRACE dataset. The obtained results show that MoAR outperforms KHM, GMM, $K$-means, W-$K$-means and SS-$K$-means. In figures 5.10 (a), (b) and (c), clustering accuracy, precision and recall of TRACE dataset are around 0.6 for $K$-means, GMM, KHM and



(a) Accuracy

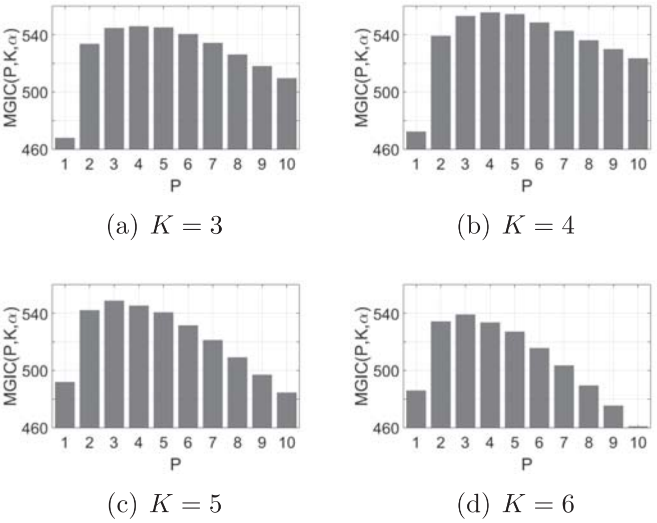
(b) Precision

(c) Recall

Figure 5.10 – Clustering performances on the TRACE dataset for $K$-means, KHM, GMM, W-$K$-means and SS-$K$-means (in the left) and MoAR for $P$ from 1 to 10 (in the right). The parameter $K$ is fixed to 4.

W-$K$-means and around 0.7 for SS-$K$-means. While, they are around 0.87 for MoAR with $P = 1$ and near to 1.0 for MoAR with $P > 1$. The obtained statistics of the clustering confirm the high performances of MoAR compared to the other approaches.

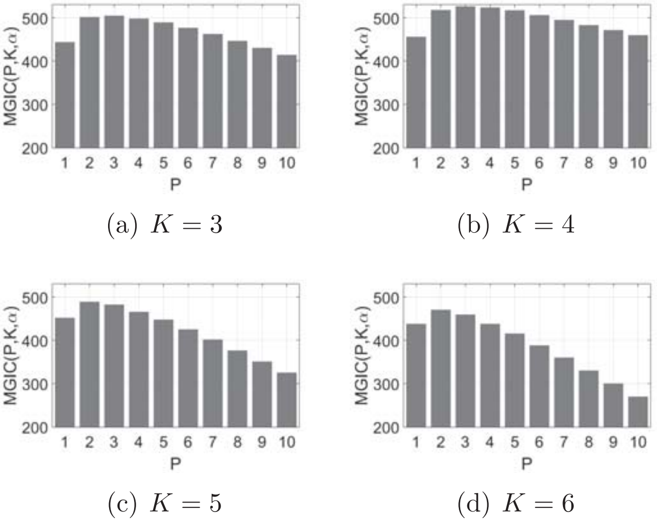
Mixture component number and AR order are selected by using MGIC($P$,$K$,$\alpha$) criterion. We study the parameter selection of our clustering approach for $\alpha = 2$ and $\alpha = 2 \ln \ln(TN)$. Figures 5.11 and 5.12 show the corresponding MGIC measures of MoAR($P$) as functions of $P$ for four different values of $K$. These figures show that MGIC values give a correct AR model order estimation for different cluster number. The highest MGIC measure corresponds to $K$ equal to 4 and $P$ between 2 and 5. These figures show that $P$ and $K$ are well estimated by MGIC criterion and confirm the results found previously in figure 5.10.



(a) $K = 3$ (b) $K = 4$

(c) $K = 5$ (d) $K = 6$

Figure 5.11 – MGIC criterion for TRACE dataset. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ and $P$. The parameter $\alpha$ is fixed to 2.



(a) $K = 3$ (b) $K = 4$

(c) $K = 5$ (d) $K = 6$

Figure 5.12 – MGIC criterion for TRACE dataset. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ and $P$. The parameter $\alpha$ is fixed to $2 \ln \ln(TN)$.

## 5.3 Surface detection of AIBO robot

AIBO (Artificial Intelligence roBOt) robot is a dog-shaped robot with four legs manufactured by Sony. It comprises several sensors, including an accelerometer with

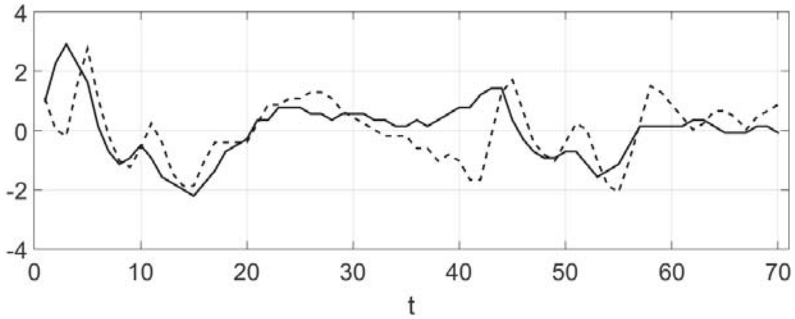
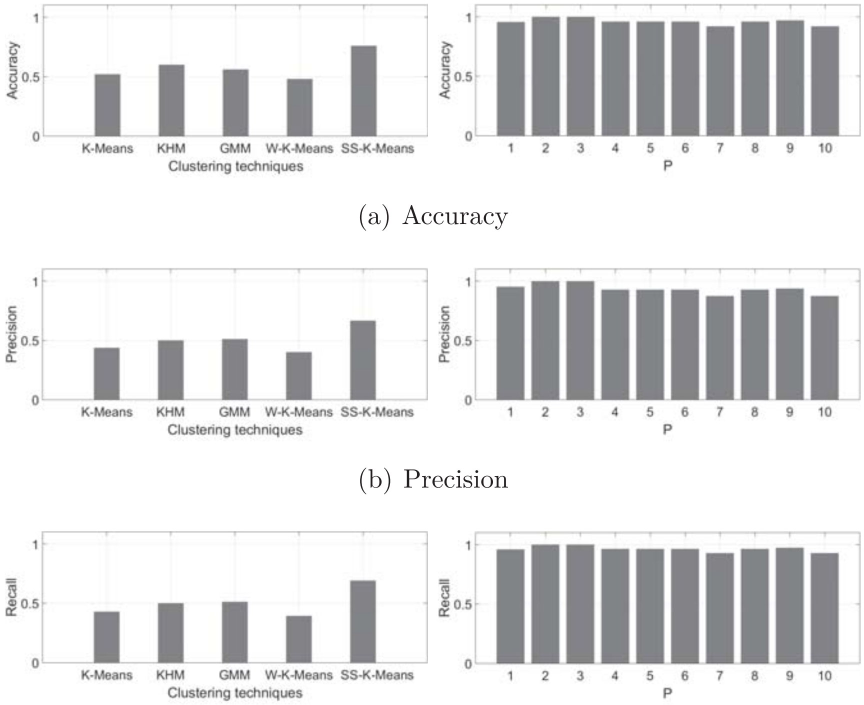


Figure 5.13 – Two time series from the RobotSurface dataset. Solid and dashed lines correspond to the robot walking respectively on carpet and cemented surfaces.

three-axis. We consider a dataset created by Vail et al. [106] and collected by an accelerometer with only the X-axis readings. In the experimental framework, AIBO robot walked on two surfaces with different nature: cement and carpet. Each instance of the dataset is a time series of 70 observations which represents one walk cycle. The dataset contains two classes of time series describing the nature of the walking surface of the robot (cement or carpet). Figure 5.13 shows an example of two time series data corresponding to the robot walking respectively on carpet and cemented surfaces. The cemented surface is considered rigid and produces more reactive forces than the carpet surface.

The robot dataset is tested by the different clustering techniques. These techniques were statistically compared in terms of clustering accuracy, precision and recall. Figure 5.14 shows the quantitative performances of the obtained clusters on the robot datasets. The obtained results show that MoAR outperforms KHM, GMM, $K$-means, W-$K$-means and SS-$K$-means. In figures 5.14 (a), (b) and (c), clustering accuracy, precision and recall of the dataset are between 0.4 and 0.5 for $K$-means and W-$K$-means, between 0.5 and 0.6 for GMM and KHM, and around 0.7 for SS-$K$-means. These measures are between 0.8 and 0.91 for MoAR with $P = 1$ and $P > 3$ and equal 1.0 for MoAR with $P = 2$ and $P = 3$. The obtained statistics of the clustering confirm the high performances of MoAR compared to the other approaches.

Parameters $K$ and $P$ are selected by using MGIC criterion for two different values of $\alpha$. Figures 5.15 and 5.16 show the corresponding MGIC measures of MoAR($P$) as functions of $P$ for four different values of $K$. These figures show that MGIC values



(a) Accuracy

(b) Precision

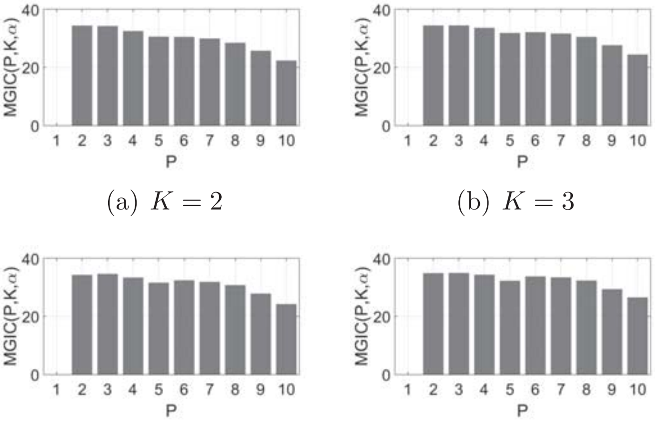
(c) Recall

Figure 5.14 – Clustering performances on RobotSurface dataset for $K$-means, KHM, GMM, W-$K$-means and SS-$K$-means (in the left) and MoAR for $P$ from 1 to 10 (in the right). The parameter $K$ is fixed to 2.

give a correct AR model order estimation for different cluster number. The highest MGIC measure corresponds to $K$ equal to 2 and $P$ between 2 and 3. The highest MGIC measure corresponds to $K$ equal to 2 and $P$ between 2 and 3. These figures show that $P$ and $K$ are well estimated by MGIC criterion and confirm the results found previously in figure 5.14.

## 5.4 Experiments on Terahertz images segmentation

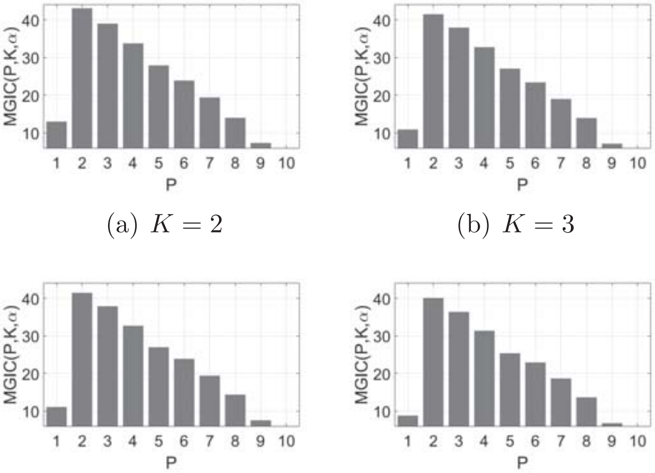
Terahertz imaging is an innovative technology of imaging [63, 112, 6, 4, 3, 42, 90]. Terahertz images can be measured by acquisition of sequences of Terahertz pulses, called time series, reflected from or transmitted through a sample. Each time series is assumed univariate and can be represented by several temporal bands behind every



(a) $K = 2$ (b) $K = 3$

(c) $K = 4$ (d) $K = 6$

Figure 5.15 – MGIC criterion for RobotSurface dataset. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ and $P$. The parameter $\alpha$ is fixed to 2.

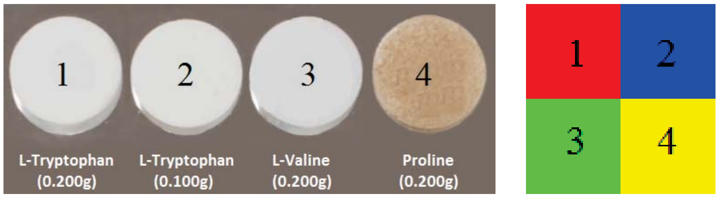


(a) $K = 2$ (b) $K = 3$

(c) $K = 4$ (d) $K = 6$

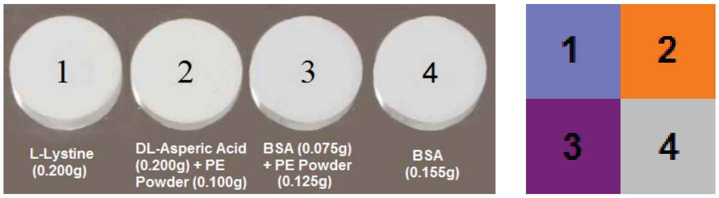
Figure 5.16 – MGIC criterion for RobotSurface dataset. MGIC($P$,$K$,$\alpha$) is presented for different values of $K$ and $P$. The parameter $\alpha$ is fixed to $2 \ln \ln(TN)$.

pixel. In this section, MoAR, $K$-means, W-$K$-means, KHM, GMM and SS-$K$-means are tested on four THz images of chemical components and cork samples. Pixels of THz images are formed respectively by 1052, 1052, 1024 and 1024 bands in the time domain. Since Terahertz images cannot be visualized (hundreds or thousands of bands), we present in figure 5.17 (a), figure 5.18 (a), figures 5.19 (a) and (b) the objects



(a) (b)

Figure 5.17 – In the left, an image of four chemical compounds acquired in visible spectrum. In the right, the ground truth of the THz image. The false colors green, blue, red and yellow correspond respectively to the chemical compounds L-Valine (0.200g), L-Tryptophan (0.100g), L-Tryptophan (0.200g) and Proline (0.200g).


1
2
3
4
L-Lystine (0.200g)
DL-Asperic Acid (0.200g) + PE Powder (0.100g)
BSA (0.075g) + PE Powder (0.125g)
BSA (0.155g)
1
2
3
4

(a) (b)

Figure 5.18 – In the left, an image of four chemical compounds acquired in visible spectrum. In the right, the ground truth of the THz image. The false colors blue, orange, purple and grey correspond respectively to the chemical compounds L-Lystine (0.200g), DL-Asperic Acid (0.200g) + PE Powder (0.100g), BSA (0.075g)+ PE Powder (0.125g) and BSA (0.155g).

acquired in the visible light for the validation. The ground truth of the chemical images, the 195$^{th}$ band and the 285$^{th}$ band of the two cork THz images are shown in figures 5.17 (b), figure 5.18 (b), figure 5.19 (c) and (d). The first chemical image comprises four compounds, L-Valine (0.200g), L-Tryptophan (0.100g), L-Tryptophan (0.200g) and Proline (0.200g), distributed into four false colored regions. The second one comprises the compounds L-Lystine (0.200g), DL-Asperic Acid (0.200g) + PE Powder (0.100g), BSA (0.075g)+ PE Powder (0.125g) and BSA (0.155g). Whereas, the third THz image shows cork matter with some cork grains as well as some voids, defects and cracks. The fourth image represents a small portion of the same cork sample with higher resolution (outlined by red frame in figure 5.19 (c)). The weights

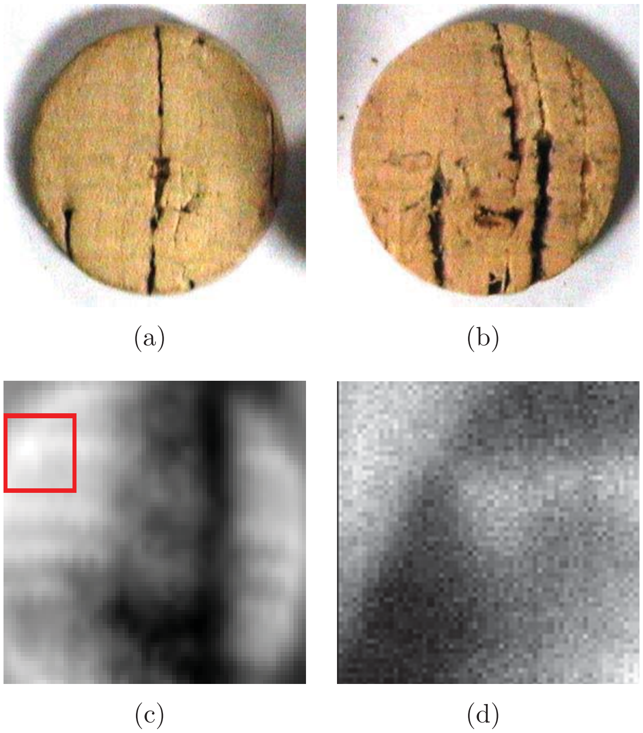


Figure 5.19 – (a) Front and (b) back images of a cork acquired in visible spectrum. (c) The $195^{th}$ band of the THz image. (d) The $195^{th}$ band of a higher resolution THz image of the portion outlined in red.

of MoAR approach and the centers of W-$K$-means, SS-$K$-means, $K$-means, KHM and GMM were initialized by random values. SS-$K$-means was used with parameters $a = 2$, $b = 2$ and three different values of parameter $c$. The segmentation of chemical and cork images was employed respectively with 4, 4, 5 and 4 clusters.

As the ground truth of the cork Terahertz images is not very precise in our work, the clustering techniques were statistically evaluated only for the chemical images. So, we have used accuracy, precision and recall measures to evaluate the results. In the case of THz images, the accuracy is the proportion of pixels correctly classified, the precision is the percentage of classified pixels which belong to the relevant class and the recall is the percentage of the relevant pixels which are correctly classified. Figure 5.20 shows the first chemical Terahertz image segmentation for the six clus-

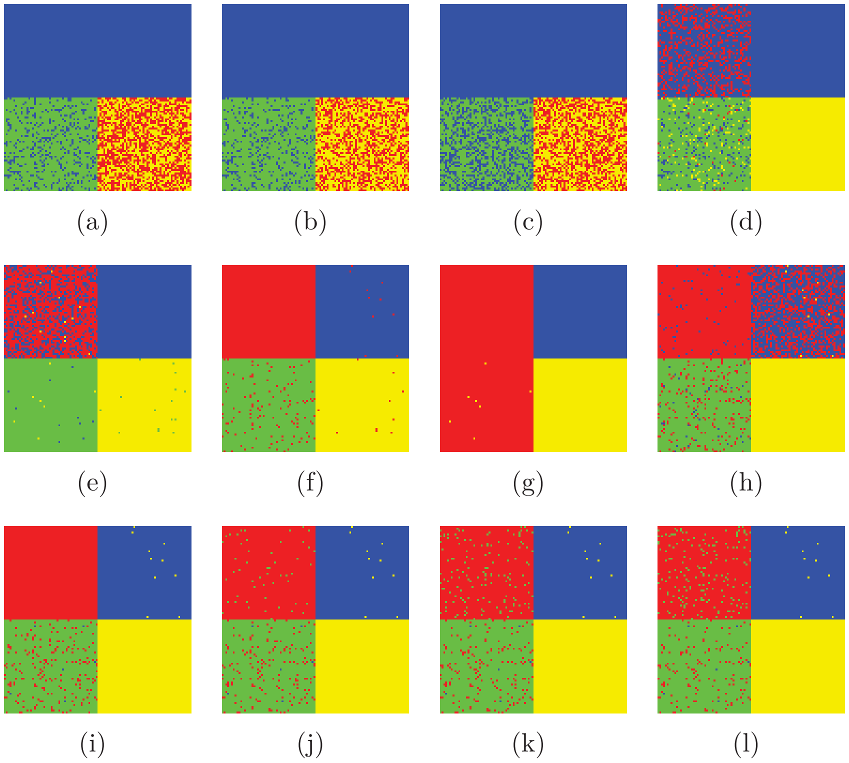
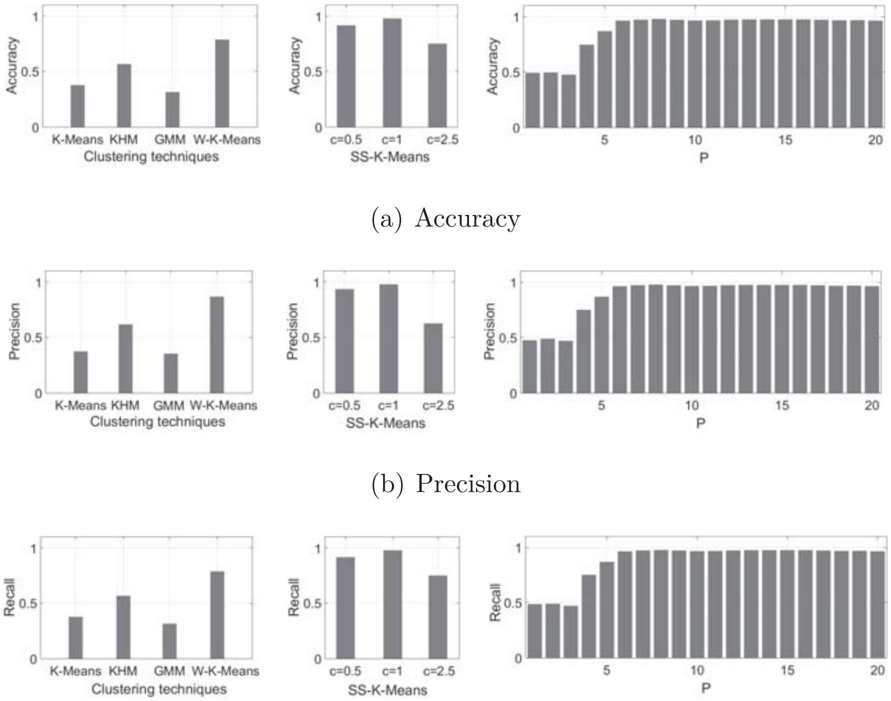


Figure 5.20 – Chemical THz image 1 segmentation for $K$-means (a), KHM (b), GMM (c), W-$K$-means (d), SS-$K$-means for $c = 0.5$ (e), $c = 1$ (f) and $c = 2.5$ (g), and MoAR for $P = 5$ (h), $P = 8$ (i), $P = 10$ (j), $P = 16$ (k) and $P = 20$ (l).

tering algorithms. In figures 5.20 (a), (b), (c), (d), (e) and (g), $K$-means, KHM, GMM, W-$K$-means and SS-$K$-means for $c = 0.5$ and $c = 2.5$ produce as output over-segmented images. L-Tryptophan (0.100g) and L-Tryptophan (0.200g) clusters are combined together in the case of $K$-means, KHM and GMM which clearly shows their segmentation shortcomings. Also, L-Valine (0.200g) and L-Tryptophan (0.200g) clusters are combined together in the case of SS-$K$-means for $c = 2.5$. L-Tryptophan (0.200g) cluster is largely affected by noisy points in the case of W-$K$-means and SS-$K$-means for $c = 0.5$. In figure 5.20 (f), SS-$K$-means for $c = 1$ shows a good segmented regions except some points of the three compounds L-Valine (0.200 g), L-Tryptophan (0.100 g) and Proline (0.200g) are misclassified. Figure 5.20 from (h) to (l) display the output regions of MoAR for different values of $P$ (5, 8, 10, 16 and 20). For $P = 5$,



(a) Accuracy

(b) Precision

(c) Recall

Figure 5.21 – Clustering performances on chemical THz image 1 for $K$-means, KHM, GMM and W-$K$-means (in the left), SS-$K$-means for different values of $c$ (in the middle), and MoAR for divers values of $P$ (in the right).

the MoAR produces as output over-segmented images. The best image segmentations are obtained when $P$ surpasses 5 which appears in figures 5.20 (i), (j), (k) and (l), the four compounds become very well identified, except some points of L-Valine (0.200g), L-Tryptophan (0.100g) and L-Tryptophan (0.200g) are misclassified. The statistics of the different approaches are shown in figure 5.21. The clustering accuracies have not surpassed the 0.4 for $K$-means and GMM, around 0.55 for KHM, 0.76 for W-$k$-means, 0.7 for SS-$K$-Means with $c = 2.5$ and surpassed 0.9 for SS-$K$-Means with $c = 0.5$ and $c = 1$. While, these measures are around 0.5 for MoAR with $P$ between 1 and 3 and increase near to 1.0 when $P$ surpass 5. The precision and the recall measures are around 0.4 for $K$-means and GMM, around 0.6 for KHM and between 0.77 and 0.85 for W-$K$-means. However, these measures are around 0.5 for MoAR with $P$ between

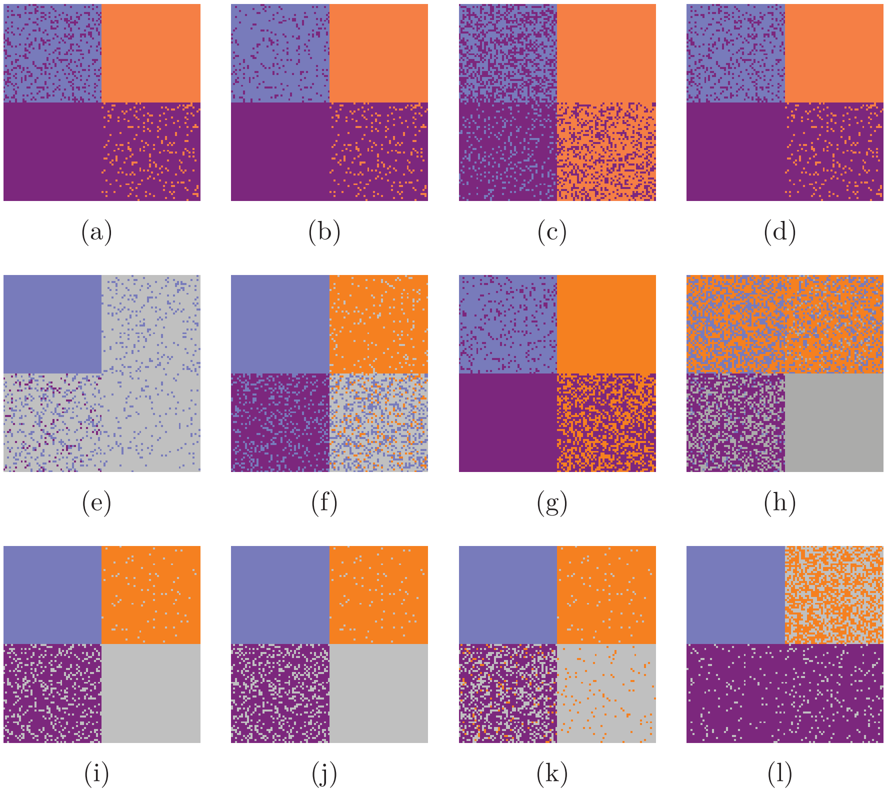
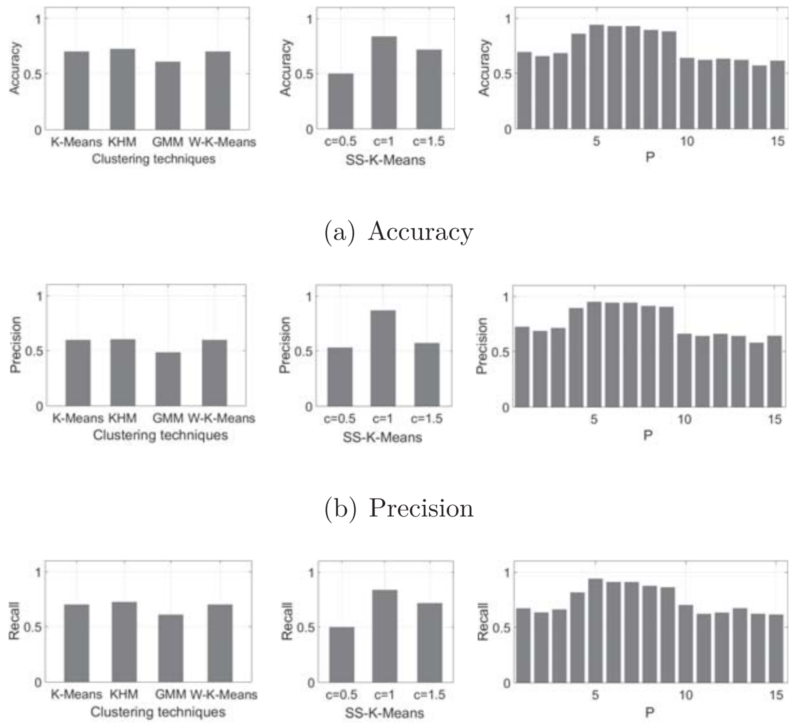


Figure 5.22 – Chemical THz image 2 segmentation for $K$-means (a), KHM (b), GMM (c), W-$K$-means (d), SS-$K$-means for $c = 0.5$ (e), $c = 1$ (f) and $c = 1.5$ (g), and MoAR for $P = 3$ (h), $P = 5$ (i), $P = 7$ (j), $P = 8$ (k) and $P = 10$ (l).

1 and 3 and become near to 1.0 when $P$ surpass 5. The obtained measures confirm the results previously illustrated and show the high performances of our approach.

Figure 5.22 shows the second chemical Terahertz image segmentation for different algorithms. In figures 5.22 (a), (b), (c), (d), (e) and (g), $K$-means, KHM, GMM, W-$K$-means and SS-$K$-means for $c = 0.5$ and $c = 1.5$ produce as output over-segmented images. BSA (0.075g)+ PE Powder (0.125g) and BSA (0.155g) clusters are combined together in the case of $K$-means, KHM and GMM which clearly shows their segmentation shortcomings. L-Tryptophan (0.200g) cluster is largely affected by noisy points in the case of W-$K$-means and SS-$K$-means for $c = 0.5$. In figure 5.22 (f), SS-$K$-means for $c = 1$ shows a good segmented regions except some points of the



(a) Accuracy

(b) Precision

(c) Recall

Figure 5.23 – Clustering performances on chemical THz image 2 for $K$-means, KHM, GMM and W-$K$-means (in the left), SS-$K$-means for different values of $c$ (in the middle), and MoAR for divers values of $P$ (in the right).

three compounds L-Valine (0.200 g), L-Tryptophan (0.100 g) and Proline (0.200g) are misclassified. Figure 5.22 from (h) to (l) display the output regions of MoAR for different values of $P$ (2, 5, 10, 13 and 15). For $P = 2$, the MoAR produces as output over-segmented images. The best image segmentations are obtained when $P$ between 3 and 9 which appears in figures 5.22 (i), (j) and (k), the four compounds become very well identified, except some points of L-Valine (0.200g) and L-Tryptophan (0.100g) are misclassified. The statistics of the different approaches are shown in figure 5.23. The clustering accuracies have not surpassed the 0.7 for $K$-means, GMM, KHM and W-$k$-means, between 0.5 and 0.7 for SS-$K$-Means with $c = 0.5$ and $c = 1.5$ and around 0.85 for SS-$K$-Means with $c = 1$. While, these measures do not surpass 0.7 for MoAR with $P$ between 1 and 3 and $P \geq 3$, and increase near to 1.0 when $P$

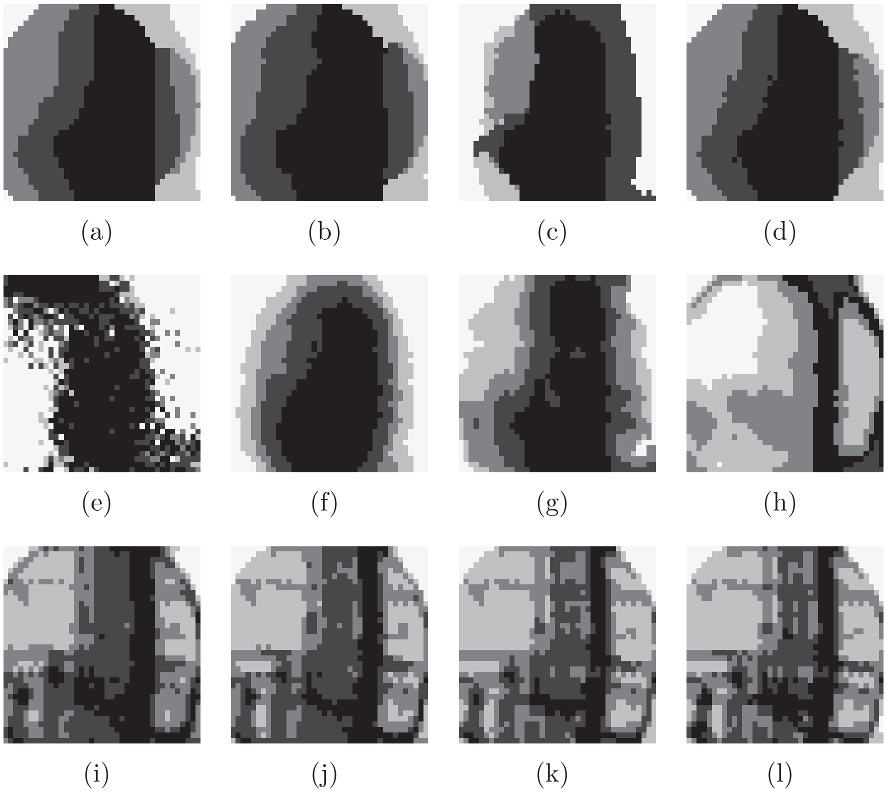


Figure 5.24 – Cork THz image 1 segmentation for $K$-means (a), KHM (b), GMM (c), W-$K$-means (d), SS-$K$-means for $c = 0.5$ (e), $c = 1$ (f) and $c = 2$ (g), and MoAR for $P = 1$ (h), 3 (i), 7 (j), 11 (k) and 15 (l).

between 4 and 9. The precision and the recall measures are between 0.5 and 0.7 for $K$-means, GMM, KHM and W-$K$-means. However, these measures are around 0.68 for MoAR with $P \leq 3$ and $P \geq 10$, and become near to 1.0 when $P$ between 4 and 9. The obtained measures confirm the results previously illustrated and show the high performances of our approach.

Figure 5.24 shows the segmentation outputs of the different clustering algorithms on the cork image 1. $K$-means, KHM, GMM, W-$K$-means and SS-$K$-means produce a wrongly segmented regions in figures 5.24 (a), (b), (c), (d), (e), (f) and (g). These techniques have not extract the details inside the cork and clearly illustrate the limitations of the five algorithms to identify defected and cracked regions. Figure 5.24

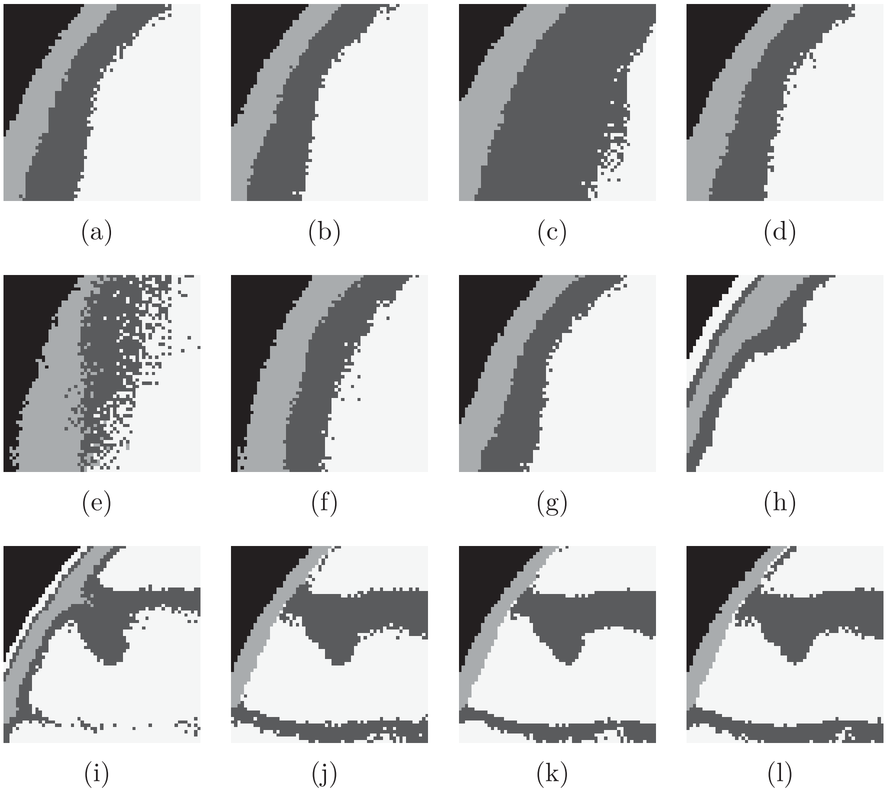
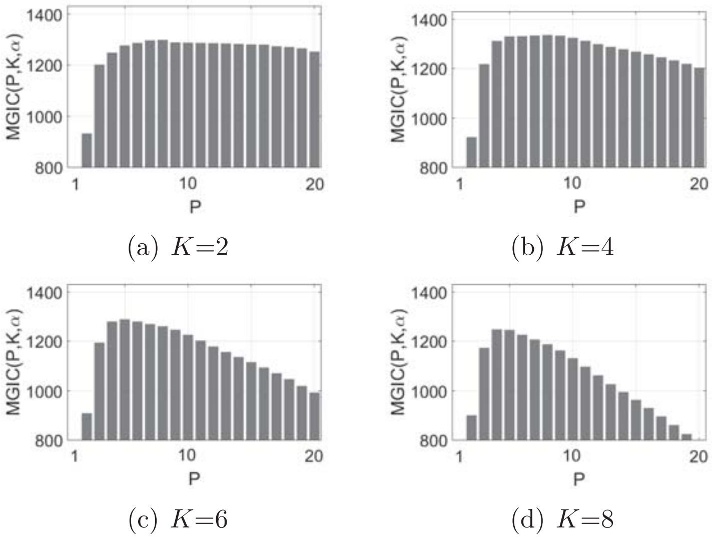


Figure 5.25 – Cork THz image 2 segmentation for $K$-means (a), KHM (b), GMM (c), W-$K$-means (d), SS-$K$-means for $c = 0.5$ (e), $c = 1$ (f) and $c = 2$ (g), and MoAR for $P = 1$ (h), 5 (i), 10 (j), 12 (k) and 15 (l).

from (h) to (l) display the obtained regions of MoAR for $P$ equal to 1, 3, 7, 11, and 15. The cork grains, the voids, the defects and the cracks are well segmented for different values of $P$.

Figure 5.25 shows the segmentation outputs of the different clustering algorithms on the cork image 2. The segmentation by $K$-means, KHM, GMM, W-$K$-means and SS-$K$-means produce a wrongly regions in figures 5.25 (a), (b), (c), (d), (e), (f) and (g). These techniques have not identify the defects and the cracks inside the cork portion and clearly illustrate their limitations compared to MoAR approach for $P$ equal to 1, 4, 8, 10, and 12.

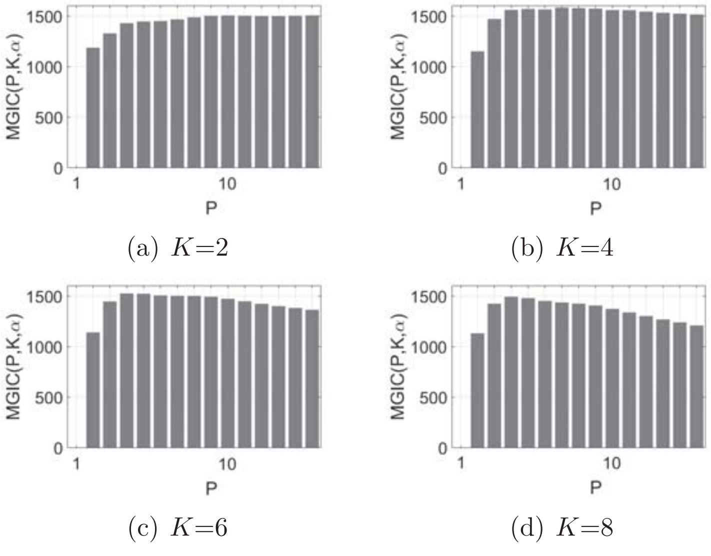
Figures 5.26, 5.27, 5.28 and 5.29 show plots of the MGIC criterion for divers



(a) $K$=2 (b) $K$=4

(c) $K$=6 (d) $K$=8

Figure 5.26 – Modified generalized information criterion (MGIC) computed after the chemical THz image 1 segmentation for different values of parameters $P$ and $K$.

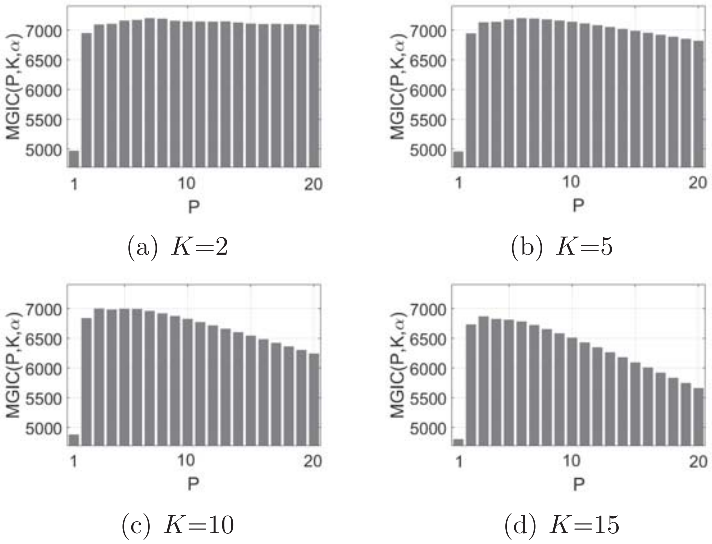


(a) $K$=2 (b) $K$=4

(c) $K$=6 (d) $K$=8

Figure 5.27 – Modified generalized information criterion (MGIC) computed after the chemical THz image 2 segmentation for different values of parameters $P$ and $K$.

orders of MoAR approach on the four THz images. The parameter $\alpha$ is fixed to $2(ln(ln(TN)))$. The best orders are around 8, 7, 6 and 5 respectively in the case



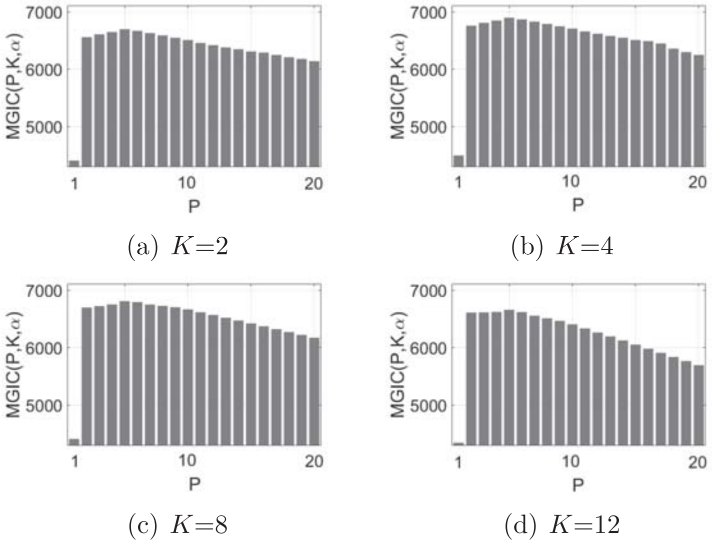
(a) $K$=2

(b) $K$=5

(c) $K$=10

(d) $K$=15

Figure 5.28 – Modified generalized information criterion (MGIC) computed after the cork THz image 1 segmentation for different values of parameters $P$ and $K$.



(a) $K$=2

(b) $K$=4

(c) $K$=8

(d) $K$=12

Figure 5.29 – Modified generalized information criterion (MGIC) computed after the cork THz image 2 segmentation for different values of parameters $P$ and $K$.

of the chemical image, the cork image and the moth image. The obtained statistics confirm the results previously illustrated in figures 5.20, 5.22, 5.24 and 5.25.

# 6 Conclusion

In this paper, we have proposed a new time series clustering approach. The MoAR approach assumes that the time series were generated by AR models and consists to recover the original autoregressive models describing each cluster of time series. The parameters of MoAR model are automatically estimated by using a modified information criterion. Our approach is tested on various artificial, transit, robotic and Terahertz datasets. Experimental results show that MoAR approach allows for interesting transient events discrimination for a safe monitoring process, successful detection of the surface nature of a mobile robot, and more accurate Terahertz image segmentation than other clustering techniques such as $K$-means, KHM, W-$K$-means, GMM and SS-$K$-means.

Our approach is shown so attractive to achieve the best clustering performances. Note that the sensitivity to initial starting conditions decreases the clustering accuracy. Furthermore, our approach deal all the features with equal importance. Feature weighting techniques can improve the accuracy of the analysis. These problems haven't been addressed in this paper and require further studies. In further work, we will deal with the initialization of the weights of AR models and the scores of the features.

# Acknowledgments

The authors would like to thank Professor John F. Federici of New Jersey Institute of Technology USA and Professor Thomas Tongue of Zomega Terahertz Corporation for the Terahertz measurements.

# Conclusion

Dans cette thèse, nous nous sommes intéressés au problème d'analyse d'images Térahertz, en proposant plusieurs approches de classification non supervisées. En résumé, trois limites ont été identifiées dans les approches de classification existantes, et pour lesquelles nous avons proposé des solutions. En premier lieu, nous avons abordé le problème d'initialisation des centres de la technique de classification $K$-means. En deuxième lieu, nous avons abordé la sélection des données à travers la pondération de caractéristiques et l'échantillonnage aléatoire statistique pour la classification des pixels. En dernier lieu, nous avons intégré la propriété de corrélation des séries chronologiques pour améliorer le processus de classification. Nos réalisations et contributions peuvent être résumées comme suit.

Dans le premier chapitre, nous avons proposé une nouvelle approche de segmentation basée sur la classification et l'échantillonnage statistique. L'approche proposée est une reformulation de la technique $K$-means dans le cadre de l'échantillon d'ensembles ordonnés pour surmonter le problème d'initialisation des centres. Des tests réalisés sur différents ensembles de données de synthèse et d'images Térahertz ont permis d'évaluer la performance de la méthode proposée par rapport à l'état de l'art. Toutefois, cette approche est face à des défis pour la sélection des caractéristiques pertinentes et le choix de la caractéristique concomitante utilisée pour trier les pixels. Dans le deuxième chapitre, nous avons abordé une stratégie de pondération de caractéristiques et une procédure d'échantillonnage aléatoire simple dans le processus de classification pour la segmentation d'images Térahertz. L'estimation automatique de la taille de l'échantillon aléatoire et le nombre de caractéristiques sélectionnées est également proposée.

Dans les deux premiers chapitres, la propriété de corrélation entre les caractéristiques de l'image Térahertz n'est pas utilisée dans le processus de classification. Dans les deux chapitres suivants, nous avons introduit une nouvelle famille de techniques de classification des séries chronologiques basées sur la régression et qui sont adaptées aux séries chronologiques. Nous avons supposé que les valeurs associées à chaque pixel d'une image Térahertz sont échantillonnées à partir d'un modèle autorégressif. La segmentation de l'image est alors vue comme un problème de classification de séries chronologiques. Ainsi, dans le troisième chapitre, la classification est formulée comme un problème d'optimisation non-linéaire. L'ordre du modèle et le nombre de classes sont automatiquement estimés en utilisant un critère de sélection de modèle. Dans le quatrième chapitre, nous avons présenté finalement une généralisation des résultats obtenus dans le troisième chapitre. Au lieu de considérer un problème de moindres carrés, nous avons proposé une approche de classification probabiliste basée sur le mélange de modèles autorégressifs. Les paramètres de l'approche proposée sont automatiquement estimés en utilisant un critère de sélection de modèle. Les résultats expérimentaux ont montré que cette approche permet de segmenter des images Térahertz avec plus de précision que d'autres approches de l'état de l'art. Cette approche est utilisée aussi pour détecter la nature de la surface d'un robot mobile et discriminer des événements transitoires pour assurer un fonctionnement sûr et économique du processus de surveillance.

Dans cette thèse, la classification des séries chronologiques a suscité un vif intérêt pour la segmentation d'images Térahertz. Comme déjà décrit dans les deux derniers chapitres, les contributions réalisées pourraient trouver des solutions à d'autres problèmes réels relevant de plusieurs domaines d'application. En effet, elles peuvent être utilisées en finance pour regrouper les sociétés cotées en bourse en examinant leurs séries chronologiques de rendements. Nous planifions l'orientation des contributions dans le domaine de la biomédecine pour étudier les signaux biologiques EEG et distinguer les séries chronologiques causées par des personnes malades ou sains.

# Bibliographie

[1] N. Abbadeni, D. Ziou, and S. Wang, *Autocovariance-based perceptual textural features corresponding to human visual perception*, Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, vol. 3, Sept 2000, pp. 901–904.

[2] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Ying Wah, *Time-series clustering – a decade review*, Information Systems **53** (2015), 16 – 38.

[3] K. Ahi, *Mathematical modeling of thz point spread function and simulation of thz imaging systems*, IEEE Transactions on Terahertz Science and Technology **PP** (2017), no. 99, 1–8.

[4] K. Ahi and M. Anwar, *Advanced terahertz techniques for quality control and counterfeit detection*, Proceedings of SPIE, vol. 9856, 2016, p. 98560G.

[5] K. Ahi, N. Asadizanjani, S. Shahbazmohamadi, M. Tehranipoor, and M. Anwar, *Terahertz characterization of electronic components and comparison of terahertz imaging with x-ray imaging techniques*, Proceedings of SPIE, vol. 9483, 2015, pp. 94830K–94830K–15.

[6] Kiarash Ahi and Mehdi Anwar, *Developing terahertz imaging equation and enhancement of the resolution of terahertz images using deconvolution*, Proceedings of SPIE, vol. 9856, 2016, pp. 98560N–98560N–18.

[7] H. Akaike, *Information theory and an extension of the maximum likelihood principle*, Second International Symposium on Information Theory (Budapest, Hungary), 1973, pp. 267–281.

[8] M.vS. Allili and D. Ziou, *An approach for dynamic combination of region and boundary information in segmentation*, International Conference Pattern Recognition, 2008, pp. 1–4.

[9] R. Cordeiro Amorim, *Learning feature weights for k-means clustering using the minkowski metric*, Ph.D. thesis, Birkbeck, University of London, 2011.

[10] C. Ari, S. Aksoy, and O. Arikan, *Maximum likelihood estimation of gaussian mixture models using stochastic search*, Pattern Recognition **45** (2012), no. 7, 2804–2816.

[11] D. H. Auston, *Picosecond optoelectronic switching and gating in silicon*, Applied Physics Letters **26** (1975), no. 3, 101–103.

[12] M. W. Ayech and D. Ziou, *K-autoregressive clustering : Application on terahertz image analysis*, International Conference Image Analysis and Recognition, Springer, 2017, pp. 145–152.

[13] M. W. Ayech and Djemel Ziou, *Terahertz image segmentation using k-means clustering based on weighted feature learning and random pixel sampling*, Neurocomputing **175, Part A** (2016), 243 – 264.

[14] M.W. Ayech and D. Ziou, *Terahertz image segmentation based on k-harmonic-means clustering and statistical feature extraction modeling*, International Conference Pattern Recognition (Tsukuba, Japan), IEEE, 2012, pp. 222–225.

[15] MW. Ayech and D. Ziou, *Segmentation of terahertz imaging using k-means clustering based on ranked set sampling*, Expert Systems with Applications **42** (2015), no. 6, 2959–2974.

[16] M.W. Ayech and Djemel Ziou, *Automated feature weighting and random pixel sampling in k-means clustering for terahertz image segmentation*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 35–40.

[17] MW. Ayech and Djemel Ziou, *Ranked k-means clustering for terahertz image segmentation*, Image Processing (ICIP), 2015 IEEE International Conference on, IEEE, 2015, pp. 4391–4395.

[18] A. J. Bagnall and G. J. Janacek, *Clustering time series from arma models with clipped data*, Proceedings of the Tenth ACM SIGKDD International Conference

on Knowledge Discovery and Data Mining (New York, NY, USA), KDD '04, ACM, 2004, pp. 49–58.

[19] T. Bardon, R. K. May, J. B. Jackson, G. Beentjes, G. de Bruin, P. F. Taday, and M. Strlič, *Contrast in terahertz images of archival documents—part i : Influence of the optical parameters from the ink and support*, Journal of Infrared, Millimeter, and Terahertz Waves **38** (2017), no. 4, 443–466.

[20] J. Bejarano, K. Bose, T. Brannan, A. Thomas, K. Adragni, N. K. Neerchal, and G. Ostrouchov, *Sampling within k-means algorithm to cluster large datasets*, Tech. Report HPCF–2011–12, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2011, (HPCF machines used : tara.).

[21] E. Berry, R. D. Boyle, A. J. Fitzgerald, and J. W. Handley, *Time frequency analysis in terahertz pulsed imaging*, Computer Vision beyond the Visible Spectrum (I. Pavlidis, ed.), Advances in Pattern Recognition, Springer Verlag, 2005, pp. 271–311.

[22] E. Berry, J. W. Handley, A. J. Fitzgerald, W. J. Merchant, R. D. Boyle, N. N. Zinovev, R. E. Miles, J. M. Chamberlain, and M. A. Smith, *Multispectral classification techniques for terahertz pulsed imaging : an example in histopathology*, Medical Engineering & Physics **26** (2004), no. 5, 423–430.

[23] M. Bessou, H. Duday, J. P. Caumes, S. Salort, B. Chassagne, A. Dautant, A. Ziéglé, and E. Abraham, *Advantage of terahertz radiation versus x-ray to detect hidden organic materials in sealed vessels*, Optics Communications **285** (2012), no. 21, 4175–4179.

[24] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum, New York, NY, 1981.

[25] N. Bouguila and D. Ziou, *Unsupervised selection of a finite dirichlet mixture model : an mml-based approach*, IEEE Transactions on Knowledge and Data Engineering **18** (2006), no. 8, 993–1009.

[26] J. W. Bowen, G. C. Walker, S. Roychowdhury, J. B. Jackson, J. F. Roberts, W. Matthews, J. Labaune, G. Mourou, M. Menu, and I. Hodder, *Image retrieval techniques for thz applications in cultural heritage*, Infrared, Millimeter, and

Terahertz Waves (IRMMW-THz), 2013 38th International Conference on, 2013, pp. 1–2.

[27] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis : forecasting and control*, John Wiley & Sons, 2015.

[28] P. S. Bradley and U. M. Fayyad, *Refining initial points for k-means clustering*, Proceedings of the Fifteenth International Conference on Machine Learning (San Francisco, CA, USA), ICML '98, Morgan Kaufmann Publishers Inc., 1998, pp. 91–99.

[29] P. J. Brockwell and R. A. Davis, *Time series : Theory and methods*, Springer-Verlag New York, Inc., New York, NY, USA, 1986.

[30] P. M. T. Broersen, *Finite sample criteria for autoregressive order selection*, IEEE Transactions on Signal Processing **48** (2000), no. 12, 3550–3558.

[31] M. A. Brun, F. Formanek, A. Yasuda, M. Sekine, N. Ando, and Y. Eishii, *Terahertz imaging applied to cancer diagnosis*, Physics in Medicine and Biology **55** (2010), no. 16, 4615–4623.

[32] R. Casarin, L. Dalla Valle, F. Leisen, et al., *Bayesian model selection for beta autoregressive processes*, Bayesian Analysis **7** (2012), no. 2, 385–410.

[33] W. L. Chan, J. Deibel, and D. M. Mittleman, *Imaging with terahertz radiation*, Reports on Progress in Physics **70** (2007), no. 8, 1325.

[34] T. W. Chen, *A study on fast k-means clustering with hierarchical data sampling for image processing*, Proceedings of Tokai-Section Joint Conference on Electrical and Related Engineering, 2010, pp. 3–4.

[35] X. Chen, Y. Ye, X. Xu, and J. Z. Huang, *A feature group weighting method for subspace clustering of high-dimensional data*, Pattern Recognition **45** (2012), no. 1, 434 – 446.

[36] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista, *The ucr time series classification archive*, July 2015, `www.cs.ucr.edu/~eamonn/time_series_data/`.

[37] Z. Chen, Z. Bai, and B. K. Sinha, *Ranked set sampling : Theory and applications*, Lecture Notes In Statistics Springer-Verlag **176** (2004).

[38] W. G. Cochran, *Sampling techniques*, third ed., John Wiley & Sons, New York, 1977.

[39] D. L. Davies and D. W. Bouldin, *Cluster separation measure*, IEEE Transactions on Pattern Analysis and Machine Intelligence **1** (1979), no. 2, 95–104.

[40] P. de Souza, *Texture recognition via autoregression*, Pattern Recognition **15** (1982), no. 6, 471 – 475.

[41] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the em algorithm*, Journal of the Royal Statistical Society, Series B (Methodological) **39** (1977), no. 1, 1–38.

[42] S. S. Dhillon, M. S. Vitiello, E. H. Linfield, A. G. Davies, M. C. Hoffmann, J. Booske, C. Paoloni, M. Gensch, P. Weightman, G. P. Williams, et al., *The 2017 terahertz science and technology roadmap*, Journal of Physics D : Applied Physics **50** (2017), no. 4, 043001.

[43] J. Dunn, *Well separated clusters and optimal fuzzy partitions*, J. Cybernet **4** (1974), 95–104.

[44] P. D'Urso, *Fuzzy clustering for data time arrays with inlier and outlier time trajectories*, IEEE Transactions on Fuzzy Systems **13** (2005), no. 5, 583–604.

[45] P. D'Urso and L. De Giovanni, *Temporal self-organizing maps for telecommunications market segmentation*, Neurocomputing **71** (2008), no. 13, 2880 – 2892, Artificial Neural Networks (ICANN 2006) / Engineering of Intelligent Systems (ICEIS 2006).

[46] P. D'Urso, L. De Giovanni, and R. Massari, *Time series clustering by a robust autoregressive metric with application to air pollution*, Chemometrics and Intelligent Laboratory Systems **141** (2015), 107 – 124.

[47] P. D'Urso, D. Di Lallo, and E. A. Maharaj, *Autoregressive model-based fuzzy clustering and its application for detecting information redundancy in air pollution monitoring networks*, Soft Computing **17** (2013), no. 1, 83–131.

[48] P. D'Urso and E. A. Maharaj, *Autocorrelation-based fuzzy clustering of time series*, Fuzzy Sets and Systems **160** (2009), no. 24, 3565 – 3589, Theme : Non-Linear Systems and Fuzzy Clustering.

[49] B. G. Quinn E. J. Hannan, *The determination of the order of an autoregression*, Journal of the Royal Statistical Society. Series B (Methodological) **41** (1979), no. 2, 190–195.

[50] L. H. Eadie, C. B. Reid, A. J. Fitzgerald, and V. P. Wallace, *Optimizing multi-dimensional terahertz imaging analysis for colon cancer diagnosis*, Expert Systems with Applications **40** (2013), no. 6, 2043–2050.

[51] B. Ferguson, *Three dimensional t-ray inspection systems*, Ph.D. thesis, University of Adelaide, School of Electrical and Electronic Engineering, 2004.

[52] B. Ferguson and X. C. Zhang, *Materials for terahertz science and technology*, Nature materials **1** (2002), no. 1, 26–33.

[53] B. Fischer, M. Hoffmann, H. Helm, G. Modjesch, and P. Uhd Jepsen, *Chemical recognition in terahertz time-domain spectroscopy and imaging*, Semiconductor Science and Technology **20** (2005), no. 7, S246–S253.

[54] A. J. Fitzgerald, E. Berry, N. N. Zinovev, G. C. Walker, M. A. Smith, and J. M. Chamberlain, *An introduction to medical imaging with coherent terahertz frequency radiation*, Physics in medicine & biology (Print) **47** (2002), no. 7, R67–R84.

[55] A. J. Fitzgerald, V. P. Wallace, M. Jimenez-Linan, L. Bobrow, R. J. Pye, A. D. Purushotham, and D. D. Arnone, *Terahertz pulsed imaging of human breast tumors*, Radiology **239** (2006), no. 2, 533–540.

[56] H. Frigui and O. Nasraoui, *Simultaneous clustering and attribute discrimination*, Fuzzy Systems, 2000. FUZZ IEEE 2000. The Ninth IEEE International Conference on, vol. 1, May 2000, pp. 158–163 vol.1.

[57] G. Gan and M. K. P. Ng, *Subspace clustering with automatic feature grouping*, Pattern Recognition **48** (2015), no. 11, 3703 – 3713.

[58] A. A. Gowen, C. O'Sullivan, and C. P. O'Donnell, *Terahertz time domain spectroscopy and imaging : emerging techniques for food process monitoring and quality control*, Trends in Food Science & Technology **25** (2012), no. 1, 40–46.

[59] R. M. Groves, B. Pradarutti, E. Kouloumpi, W. Osten, and G. Notni, *2d and 3d non-destructive evaluation of a wooden panel painting using shearography and terahertz imaging*, Ndt & E International **42** (2009), no. 6, 543–549.

[60] J. P. Guillet, B. Recur, L. Frederique, B. Bousquet, L. Canioni, I. Manek-Hönninger, P. Desbarats, and P. Mounaix, *Review of terahertz tomography techniques*, Journal of Infrared, Millimeter, and Terahertz Waves **35** (2014), no. 4, 382–411.

[61] J. W. Handley, *Time frequency analysis techniques in terahertz pulsed imaging*, Ph.D. thesis, University of Leeds, 2003.

[62] P. G. P. Ho, *Image segmentation by autoregressive time series model*, Image Segmentation, InTech, 2011.

[63] B. B. Hu and M. C. Nuss, *Imaging with terahertz waves*, Optics Letters **20** (1995), no. 16, 1716–1718.

[64] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, *Automated variable weighting in k-means type clustering*, IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005), no. 5, 657–668.

[65] K. Ikushima and S. Komiyama, *Imaging by terahertz photon counting*, Comptes Rendus Physique **11** (2010), 444–456.

[66] J. B. Jackson, J. Bowen, G. Walker, J. Labaune, G. Mourou, M. Menu, and K. Fukunaga, *A survey of terahertz applications in cultural heritage conservation science*, IEEE Transactions on Terahertz Science and Technology **1** (2011), no. 1, 220–231.

[67] A. K. Jain, *Data clustering : 50 years beyond k-means*, Pattern Recognition Letters **31** (2010), no. 8, 651 – 666, Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

[68] A. K. Jain, M. N. Murty, and P. J. Flynn, *Data clustering : A review*, ACM Computer Survey **31** (1999), no. 3, 264–323.

[69] N. Kamba, Y. Tsuchiya, A. Okimoto, and K. Fukunaga, *Internal structure observation of a japanese panel painted screen by terahertz imaging technique*, Proceeding 10th International Conference on non-destructive investigations and microanalysis for the diagnostics and conservation of cultural and environmental heritage, NDT-41, 2011.

[70] D. D. W. Karl and A. D. Wilson, *Infrastructs : Fabricating information inside physical objects for imaging in the terahertz region*, ACM Transactions on Graphics **32** (2013), no. 4, 138 :1–138 :10.

[71] M. Kowalski, N. Palka, M. Piszczek, and M. Szustakowski, *Hidden object detection system based on fusion of thz and vis images*, Acta Physica Polonica A **124** (2013), no. 3, 490–493.

[72] E. Leiss-Holzinger, K. Wiesauer, H. Stephani, D. Stifter B. Heise, and B. Kriechbaumer, *Imaging of the inner structure of cave bear teeth by novel non-destructive techniques*, Palaeontologia electronica **18** (2015), no. 18.1.1T, 1–15.

[73] T. Warren Liao, *Clustering of time series data-a survey*, Pattern Recognation **38** (2005), no. 11, 1857–1874.

[74] J. Lin, M. Vlachos, E. J. Keogh, and D. Gunopulos, *Iterative incremental clustering of time series*, EDBT, 2004.

[75] J. Liping, M. K. Ng, and J. Z. Huang, *An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data*, Knowledge and Data Engineering, IEEE Transactions on **19** (2007), no. 8, 1026–1041.

[76] H. B. Liu, H. Zhong, N. Karpowicz, Y. Chen, and X. C. Zhang, *Terahertz spectroscopy and imaging for defense and security applications*, Proceedings of the IEEE **95** (2007), no. 8, 1514–1527.

[77] J. B. MacQueen, *Some methods for classification and analysis of multivariate observations*, Proceeding of the fifth Berkeley Symposium on Mathematical Statistics and Probability (L. M. Le Cam and J. Neyman, eds.), vol. 1, University of California Press, 1967, pp. 281–297.

[78] E. A. Maharaj, *Clusters of time series*, Journal of Classification **17** (2000), no. 2, 297–314.

[79] E. A. Maharaj and P. D'Urso, *Fuzzy clustering of time series in the frequency domain*, Information Sciences **181** (2011), no. 7, 1187 – 1211.

[80] G. A. McIntyre, *A method of unbiased selective sampling, using ranked sets*, Australian Journal of Agricultural Research **3** (1952), 385–390.

[81] G. McLachlan and D. Peel, *Finite mixture models*, Wiley Series in Probability and Statistics, 2000.

[82] D. M. Mittleman, M. Gupta, R. Neelamani, R. G. Baraniuk, J. V. Rudd, and M. Koch, *Recent advances in terahertz imaging*, Applied Physics B : Lasers and Optics **68** (1999), 1085–1094.

[83] D. M. Mittleman, R. H. Jacobsen, and M. C. Nuss, *T-ray imaging*, IEEE Journal on Selected Topics in Quantum Electronics **2** (1996), 679–692.

[84] C. S. Moller-Levet, F. Klawonn, K. H. Cho, and O. Wolkenhauer, *Fuzzy clustering of short time-series and unevenly distributed sampling points*, Advances in Intelligent Data Analysis V (Berlin, Heidelberg) (M. R. Berthold, H. J. Lenz, E. Bradley, R. Kruse, and C. Borgelt, eds.), Springer Berlin Heidelberg, 2003, pp. 330–340.

[85] S. Nakajima, H. Hoshina, M. Yamashita, C. Otani, and N. Miyoshi, *Terahertz imaging diagnostics of cancer tissues with a chemometrics technique*, Applied Physics Letters **90** (2007), no. 4.

[86] M. Nazari, J. Shanbehzadeh, and A. Sarrafzadeh, *Fuzzy-c-means based on automated variable feature weigthing*, Proceedings of the International MultiConference of Engineers and Computer Scientists, 2013.

[87] U. Orhan, M. Hekim, and M. Ozer, *Eeg signals classification using the k-means clustering and a multilayer perceptron neural network model*, Expert Systems with Applications **38** (2011), no. 10, 13475 – 13481.

[88] G. P. Patil, *Ranked set sampling*, In Encyclopedia of Environmetrics **3** (2002), 1684–1690.

[89] G. P. Patil, A. K. Sinha, and C. Taillie, *Ranked set sampling for multiple characteristics*, International Journal of Ecology and Environmental Sciences **20** (1994), 94–109.

[90] A. Y. Pawar, D. D. Sonawane, K. B. Erande, and D. V. Derle, *Terahertz technology and its applications*, Drug invention today **5** (2013), no. 2, 157–163.

[91] W. D. Penny and S. J. Roberts, *Variational bayes for non-gaussian autoregressive models*, Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop (Cat. No.00TH8501), vol. 1, Dec 2000, pp. 135–144 vol.1.

[92] R. J. Povinelli, M. T. Johnson, A. C. Lindgren, and J. Ye, *Time series classification using gaussian mixture models of reconstructed phase spaces*, IEEE Transactions on Knowledge and Data Engineering **16** (2004), no. 6, 779–783.

[93] D. M. W. Powers, *Evaluation : from precision, recall and f-measure to roc, informedness, markedness and correlation*, International Journal of Machine Learning Technology **2** (2011), no. 1, 37–63.

[94] R. A. Redner and H. F. Walker, *Mixture densities, maximum likelihood and the em algorithm*, SIAM Review **26** (1984), no. 2, 195–239.

[95] A. Redo-Sanchez, *The terahertz wave ebook*, Tech. report, Zomega Terahertz Corporation, 2012.

[96] P. J. Rousseeuw, *Silhouettes : A graphic aid to the interpretation and validation of cluster analysis*, Journal of Computational and Applied Mathematics **20** (1987), no. 1, 53—65.

[97] D. Roverso, *Multivariate temporal classification by windowed wavelet decomposition and recurrent neural networks*, 3rd ANS international topical meeting on nuclear plant instrumentation, control and human-machine interface, vol. 20, 2000.

[98] G. Schwarz, *Estimating the dimension of a model*, The Annals of Statistics **6** (1978), no. 2, 461–464.

[99] V. Sharma, D. Arya, and M. Jhildiyal, *Terahertz technology and its applications*, IEEE International Conference on Advanced Computing & Communication Technologies (ICACCT), 2011, pp. 175–178.

[100] Siuly, X. X. Yin, S. Hadjiloucas, and Y. Zhang, *Classification of thz pulse signals using two-dimensional cross-correlation feature extraction and non-linear classifiers*, Computer Methods and Programs in Biomedicine **127** (2016), 64 – 82.

[101] H. Stephani, *Automatic segmentation and clustering of spectral terahertz data*, Ph.D. thesis, 2012.

[102] H. Stephani, J. Jonuscheit, C. Robine, and B. Heise, *Automatically detecting peaks in terahertz time-domain spectroscopy*, Pattern Recognition (ICPR), 2010 20th International Conference on, IEEE, 2010, pp. 4468–4471.

[103] K. Takahasi and K. Wakimoto, *On unbiased estimates of the population mean based on the sample stratified by means of ordering*, Annals of the Institute of Statistical Mathematics **20** (1968), no. 1, 1–31.

[104] M. E. Thompsom, *Theory of sample surveys*, New York Chapman and Hall, 1997.

[105] K. S. Tuncel and M. G. Baydogan, *Autoregressive forests for multivariate time series modeling*, Pattern Recognition **73** (2018), no. Supplement C, 202 – 215.

[106] D. Vail and M. Veloso, *Learning from accelerometer data on a legged robot*, IFAC Proceedings Volumes **37** (2004), no. 8, 822 – 827, IFAC/EURON Symposium on Intelligent Autonomous Vehicles, Lisbon, Portugal, 5-7 July 2004.

[107] Y. G. Wang, Y. Ye, and D. A. Milton, *Efficient designs for sampling and subsampling in fisheries research based on ranked sets*, ICES Journal of Marine Science **66** (2009), 928–934.

[108] X. X. Yin, S. Hadjiloucas, Y. Zhang, M. Y. Su, Y. M., and D. Abbott, *Pattern identification of biomedical images with time series : Contrasting {THz} pulse imaging with dce-mris*, Artificial Intelligence in Medicine **67** (2016), 1 – 23.

[109] X. X. Yin, B. W. H. Ng, B. Ferguson, S. P. Mickan, and D. Abbott, *Statistical model for the classification of the wavelet transforms of t-ray pulses*, 18th International Conference Pattern Recognition, vol. 3, 2006, pp. 236–239.

[110] X. X. Yin, B. W. H. Ng, B. M. Fischer, B. Ferguson, and D. Abbott, *Support vector machine applications in terahertz pulsed signals feature sets*, IEEE Sensors Journal **7** (2007), no. 12, 1597–1608.

[111] B. Zhang, M. Hsu, and U. Dayal, *K-harmonic means - a data clustering algorithm*, Tech. Report HPL-I999-I24, Hewlett-Packard Laboratories, 1999.

[112] X. C. Zhang, *Terahertz wave imaging : horizons and hurdles*, Physics in Medicine and Biology **47** (2002), no. 21, 3667–3677.

[113] H. Zhong, A. Redo-Sanchez, and X. C. Zhang, *Identification and classification of chemicals using terahertz reflective spectroscopic focal-plane imaging system*, Optics Express **14** (2006), no. 20, 9130–9141.

[114] B. Zhu, Y. Chen, K. Deng, W. Hu, and Z. S. Yao, *Terahertz science and technology and applications*, PIERS Proc., Beijing (2009), 1166.