Les représentations et l'analyse statistique des signaux

Michel Barret

Mémoire d'habilitation à diriger des recherches

de l'UNIVERSITÉ PARIS-SUD XI

présenté par

Michel BARRET

Équipe Information, Multimodalité & Signal
SUPELEC

SUR LES REPRÉSENTATIONS ET L'ANALYSE STATISTIQUE DES SIGNAUX

Soutenu le 4 juin 2010 devant le jury composé de :

M. Pierre DUHAMEL	**Président**
M. Marc ANTONINI	**Rapporteur**
M. Christian JUTTEN	**Rapporteur**
M. Jean-Christophe PESQUET	**Rapporteur**
M. Jacques OKSMAN	
M. Dinh-Tuan PHAM	

Table des matières

Chapitre 1

Introduction

Dans l'introduction de son livre *Extrapolation, interpolation and smoothing of stationary time series*, Wiener écrivait :

> « This book represents an attempt to unite the theory and practice of two fields of work which are of vital importance in the present emergency, and which have a complete natural methodological unity, but which have up to the present drawn their inspiration from two entirely distinct traditions, and which are widely different in their vocabulary and the training of their personnel. These two fields are those of time series in statistics and of communication engineering. »

Ainsi, dès son origine, le traitement statistique du signal se situe au carrefour des mathématiques et des sciences pour l'ingénieur. C'est probablement pour cela que ce domaine n'a cessé de me passionner depuis que je l'ai découvert avec Bernard Picinbono il y a une vingtaine d'années [12].

L'objet de ce mémoire est de présenter les principales contributions que j'ai pu apporter en traitement statistique du signal grâce à diverses collaborations. Une première partie porte sur la stabilité des filtres et des systèmes linéaires, en particulier celle des filtres numériques bi-dimensionnels. Une deuxième partie expose des résultats en compression d'images, en particulier sur les espaces de représentations bien adaptés à leur codage.

En suivant Shannon, les pionniers de la théorie du codage ont délimité un cadre mathématique très général, distinguant la compression sans perte de celle avec pertes. Pour cette dernière, ils ont étudié en détails les propriétés qui découlaient de l'hypothèse de données gaussiennes. Les méthodes développées par les ingénieurs, qui sont basées en général sur des résultats théoriques utilisant cette hypothèse, n'ont cessé de s'améliorer tout en offrant de plus en plus de souplesse aux utilisateurs, approchant aujourd'hui des limites théoriques. Notre objectif est d'étendre quelques propriétés mathématiques utiles aux ingénieurs en communication en relâchant l'hypothèse de données gaussiennes.

La suite de l'introduction donne un résumé des différents chapitres du mémoire et une cohérence entre eux. Elle se termine par une présentation rapide de mes activités d'enseignement et des contrats industriels auxquels j'ai participé.

L'étude de la stabilité des filtres récursifs bi-dimensionnels est le sujet de ma thèse [48], dirigée par Messaoud Benidir et soutenue en décembre 1993. Les publications relatives à ma thèse sont les suivantes : [10, 11, 31, 32, 33, 34, 35] (2 revues et 5 congrès). J'ai également été invité à trois séminaires mathématiques [54, 53, 52] pour présenter les résultats de ma thèse.

Après ma thèse, j'ai continué à travailler sur la stabilité des filtres récursifs 2D (voire ND), aussi bien sous l'angle historique de l'étude de la localisation des zéros d'un polynôme que sous celui de l'efficacité des algorithmes de test de stabilité, en particulier de leur robustesse face aux erreurs d'arrondis. Ces recherches ont donné lieu aux publications suivantes : [1, 2, 3, 8, 9, 47] (1 livre, 2

chapitres de livre traduits en anglais, 2 revues et 1 pré-publication). J'ai évalué et critiqué plus d'une dizaine d'articles sur le sujet soumis pour publication dans diverses revues internationales.

Les filtres récursifs mono-dimensionnels trouvaient dans les années 80 de nombreuses applications, par exemple en codage de la parole avec le codage par prédiction linéaire, grâce à la théorie de la prédiction à passé fini ou infini (voir par exemple le livre de cours [55]) et à la représentation en treillis des filtres récursifs, qui permet un très bon contrôle des erreurs d'arrondi. Dans le cas bi-dimensionnel, les filtres récursifs trouvaient très peu d'applications. À l'époque de ma thèse cela pouvait être dû aux difficultés pratiques que l'on rencontrait pour tester leur stabilité. Mais dans la décennie qui a suivi, ils n'ont pas trouvé beaucoup d'applications nouvelles. Les bancs de filtres et les décompositions en ondelettes ont eu beaucoup plus de succès.

Dès la fin des années 80, de nouvelles représentations des signaux, issues des décompositions en ondelettes de Mallat, Meyer et Daubechies, ou des *lapped transforms* de Malvar ou encore des bancs de filtres à reconstruction parfaite de Smith & Barnwell, Vaidyanathan, Vetterli, ..., commencent à être appliquées pour du codage d'images. Elles remplaceront le codage par transformée (transformée en cosinus discrète de JPEG) qui a le défaut, aux faibles débits, de rendre visible le découpage par blocs (*blocking effect*) de l'image. En 1992, Antonini, Barlaud, Mathieu et Daubechies (*IEEE Trans. Image Processing*, **IP-1**, 2, 1992) publient un article de référence sur l'emploi des ondelettes à support compact pour du codage d'images et montrent les bonnes performances de l'ondelette dite 9/7 de Daubechies. Au milieu des années 90, Sweldens introduit la représentation en *lifting scheme* des ondelettes bi-orthogonales (*Proc. SPIE*, **2569**, septembre 1995) et montre avec Daubechies (*J. Fourier Anal. Appl.*, **4**, 3, 1998) que toute décomposition en ondelettes bi-orthogonale utilisant des filtres RIF admet une représentation en *lifting scheme*. Cette représentation permet de construire très facilement des transformations réversibles (d'entiers en entiers) approximant les décompositions en ondelettes (*integer to integer wavelets*) et donnant de bonnes performances en codage sans perte des images (voir l'article de Calderbank, Daubechies, Sweldens et Yeo, *Applied and Computational Harmonic Analysis*, **5**, 3, 1998). Le *lifting scheme* associe décomposition multi-résolution et prédiction[1]. Il est alors naturel d'étudier des structures en *lifting scheme* dans lesquelles les étapes de prédiction à coefficients fixes sont remplacées par des étapes d'estimations linéaires en moyenne quadratique. À la fin des années 90, pour du codage avec pertes et aussi sans perte des images, Gerek et Çetin (*IEEE Trans. on Image Processing*, **9**, 10, 2000) ont appliqué l'algorithme du gradient stochastique pour que l'étape dite de prédiction de la structure en *lifting scheme* s'adapte au signal à coder.

En utilisant le vocabulaire de la théorie de l'estimation, Gerek et Çetin n'ont étudié que l'estimation linéaire adaptative, estimant la valeur d'un signal de sous-bande, disons J_2 à la position (m, n) comme combinaison linéaire des échantillons du signal de l'autre sous-bande (J_1) uniquement, ils n'ont pas utilisé l'information disponible portée par les valeurs du signal J_2 dans un voisinage causal ou semi-causal de la position (m, n). En d'autres termes, avec le vocabulaire de la théorie de l'estimation, ils ont appliqué l'estimation linéaire seule, sans l'associer à de la prédiction linéaire et en termes d'informations, ils n'ont pas utilisé toute l'information disponible pour adapter l'étape de prédiction de la structure en *lifting scheme* au signal à coder. C'est ce que nous avons voulu faire au début des années 2000 en encadrant (80% d'encadrement) la thèse de H. Bekkouche, dirigée par J. Oksman (20% d'encadrement), sur la synthèse de bancs de filtres adaptés avec des applications en codage

[1] Attention au vocabulaire, qui n'a pas le même sens en théorie de la décision et en filtrage numérique. Dans une structure en *lifting scheme* il est d'usage d'appeler "prédiction" l'étape d'estimation : après avoir partagé le signal original en deux signaux de sous-bandes, l'étape dite de prédiction consiste à prédire la valeur d'un de ces signaux comme combinaison linéaire des échantillons de l'**autre** signal, mais dans la théorie de l'estimation, le terme "prédiction" est réservé au cas où la valeur d'un signal à l'instant n (ou la position (m, n)) est estimée à partir d'un voisinage causal du **même** signal.

sans perte des images. En recherchant pour le banc d'analyse des filtres prédicteurs bi-dimensionnels optimaux, les filtres du banc de synthèse sont récursifs bi-dimensionnels. Nous espérions ainsi trouver une application de ces filtres dans le cas non séparable (c'est-à-dire quand ils ne se réduisent pas à des filtres mono-dimensionnels). Ces travaux ont fait l'objet des publications suivantes :

- un article long dans une revue internationale [6] ;
- quatre congrès internationaux avec actes et comité de lecture [27, 28, 29, 30] ;
- un congrès national avec actes, sans comité de lecture [41].

Après avoir rencontré des difficultés personnelles qui l'ont retardé, Hocine Bekkouche a soutenu sa thèse de l'Université Paris-Sud, spécialité Traitement du Signal, intitulée "Synthèse de bancs de filtres adaptés, application à la compression des images", le 18 juin 2007.

Grâce aux travaux réalisés avec H. Bekkouche et J. Oksman, j'ai pu approfondir mes connaissances sur les bancs de filtres à reconstruction parfaite et le codage sans perte des images. L'étude des bancs de filtres à reconstruction parfaite a également donné lieu à la rédaction d'un polycopié de Supélec sur le sujet [57]. De plus, en collaboration avec J.-L. Collette, J. Oksman, P. Duhamel et d'autres membres de l'équipe de recherche Signaux et Systèmes Électroniques de Supélec (EA-2523), j'ai participé à l'étude des bancs de filtres hybrides pour la conversion analogique/numérique très large bande et j'ai contribué à la rédaction de trois articles présentés dans des congrès internationaux à comité de lecture ([19, 24, 25]).

En 2000, le nouveau standard en codage d'images, JPEG2000 part 1, fixait les règles de codage avec pertes ou quasi sans perte pour les images naturelles. À cette époque, nous avons choisi de travailler sur la compression sans perte en cherchant des partenaires utilisateurs d'un tel codage en imagerie satellitaire ou médicale. Nous avons alors pris contact avec Catherine Lambert Nebout, ingénieur docteur du CNES.

Comme j'enseignais à des élèves de troisième année de Supélec le cours de traitement statistique du signal [55], pour réduire l'écart entre mon enseignement et mes recherches, je souhaitais poursuivre ces dernières en les orientant vers l'analyse statistique des données multi-dimensionnelles (2-D essentiellement) non gaussiennes.

Or, depuis un peu plus d'une décennie, la séparation aveugle de sources (*blind source separation*), puis l'analyse en composantes indépendantes étaient apparues et s'étaient développées, offrant divers algorithmes efficaces dans plusieurs domaines d'applications, comme les télécommunications. De plus, en analyse multispectrale, Nuzillard et Bijaoui avaient appliqué différentes méthodes de séparation de sources sur quatre images de la galaxie 3C120 prises par le télescope Hubble (D. Nuzillard et A. Bijaoui, *Astronomy and Astrophysics*, septembre 2000). Ils avaient montré qu'avec une interprétation subjective des résultats fournis par des algorithmes de séparation de sources et en modifiant ces derniers (en particulier en imposant aux matrices de mélange d'avoir des coefficients positifs), on pouvait obtenir quatre images très différentes entre elles à l'œil (contrairement aux résultats de l'analyse en composantes principales ou des décompositions en ondelettes). De plus, les quatre images qu'ils avaient obtenues représentaient des sources différentes, qui avaient toutes un sens en astrophysique. Ce qui n'est pas très étonnant, car c'était le critère d'arrêt choisi par Nuzillard et Bijaoui dans leur démarche itérative d'adaptation des algorithmes de séparation de sources et d'interprétation des résultats.

Enfin, le caractère non gaussien des coefficients d'ondelettes associés aux images naturelles, l'efficacité d'algorithmes de codage exploitant la redondance entre coefficients d'ondelettes inter-bandes (SPIHT de Said et Pearlman, *IEEE Trans. Circuits and Systems for Video Technology*, **6**, 3, 1996) ou intra-bande (EBCOT de Taubman, *IEEE Trans. Image Processing*, **IP-9**, 7, 2000)—malgré la faible corrélation statistique entre ces coefficients—, et la publication dans des revues de traitement d'images d'analyses de l'information mutuelle entre coefficients d'ondelettes (Liu et Moulin, *IEEE*

Trans. Image Processing, **10**, 11, 2001) ont été à l'origine du sujet de thèse de M. Narozny : *Analyse en composantes indépendantes et compression de données*. C'est une thèse de l'Université Paris-Sud en spécialité du traitement du signal, pour laquelle j'ai été directeur, grâce à une dérogation me dispensant de l'habilitation à diriger des recherches. J'ai réalisé 80% de l'encadrement, J. Oksman et P. Duhamel en ayant chacun réalisé 10% également. La thèse a été soutenue le 12 décembre 2005. Elle a donné lieu aux publications suivantes :

- un article long dans une revue internationale [7] ;
- quatre congrès internationaux avec actes et comité de lecture [18, 22, 23, 26] ;
- trois ateliers avec actes sans comité de lecture [40, 39, 37] ;
- et deux rapports de recherche [45, 46].

Avec M. Narozny, nous avons commencé par étudier l'ajout, dans une structure en *lifting scheme*, d'une étape réduisant l'information mutuelle entre sous-bandes avec des applications en codage sans perte d'images en niveaux de gris ou en couleur. Cela a fait l'objet d'une publication dans un congrès international avec comité de lecture [26], où nous avons rencontré Dinh-Tuan Pham. Grâce au soutien de l'Action Concertée Incitative ACI^{2}M décrite ci-dessous, nous avons étudié le codage par transformée et la quantification scalaire ou vectorielle pour généraliser la notion de gain de codage d'une transformation sans l'hypothèse de données gaussiennes. Cela a fait l'objet de deux rapports de recherche ([45, 46]) et d'une présentation des avancées du projet aux Journées ACI Masse de données des 16 et 17 septembre 2004 [40], et aux Journées Pari-STIC des années 2005 et 2006 [39, 37]. À partir de l'expression du gain de codage généralisé, nous avons modifié l'algorithme `ICAinf` (D. T. Pham, *IEEE Trans. Signal Processing*, **52**, 10, 2004) pour construire deux algorithmes maximisant le gain de codage, l'un (`OrthICA` pour *Orthogonal Independent Component Analysis*) en imposant à la matrice de séparation d'être orthogonale et l'autre (`GCGsup` pour *Generalized Coding Gain Supremum*) sans cette contrainte. Appliqués à des signaux synthétiques, décorrélés et non gaussiens, ces algorithmes donnent des gains de codage importants par rapport à ceux obtenus avec la transformation de Karhunen-Loève[2]. Ces travaux ont fait l'objet de publications dans trois congrès internationaux avec actes [18, 22, 23], et dans une revue internationale [7].

Dans le cadre de l'Action Concertée Incitative (ACI) Masse de Données du Ministère de l'Éducation Nationale et de la Recherche, suite aux contacts pris avec Catherine Lambert Nebout du CNES, et grâce—entre autres—aux soutiens de Pierre Duhamel et de Dinh-Tuan Pham, nous avons soumis en 2003 un projet intitulé *Analyse en composantes indépendantes appliquée au codage d'images multicomposantes*, avec quatre laboratoires partenaires :

- le CNES, centre spatial de Toulouse, avec C. Lambert Nebout (remplacée en cours de projet par C. Thiebaut),
- l'I3S (Informatique, Signaux et Systèmes de Sophia-Antipolis), avec P. Comon,
- le LMC (Lab. de Modélisation et Calcul) de l'IMAG (Grenoble), devenu aujourd'hui le Laboratoire Jean Kuntzmann, avec D. T. Pham,
- le L2S (Lab. des Signaux et Systèmes) (Gif-sur-Yvette), avec P. Duhamel et A. Mohammad-Djafari,

qui a été accepté. C'est le projet ACI^{2}M, dont j'ai été le coordinateur de 2003 à 2006. Voici une description courte du projet :

> « Les quantités de données transmises ou stockées en imagerie satellitale sont considérables. Nous étudions le problème du codage d'une famille d'images (appelée image multi-composante). Un codage progressif et pouvant aller jusqu'au sans perte trouve des

[2]C'est le nom utilisé en codage d'images pour désigner le changement de repère de l'analyse en composantes principales.

applications potentielles, aujourd'hui, en imagerie multi-composante satellitale ou médicale. Pour des images fixes, les décompositions multi-résolutions réversibles sont bien adaptées à des codeurs entropiques par arbre de zéros (du type SPIHT) et permettent un codage progressif et sans perte. Par ailleurs, l'analyse en composantes indépendantes (ACI) a atteint un grand degré de maturité théorique et des algorithmes performants, qui couvrent un large éventail de situations, sont disponibles. Il est généralement admis que les algorithmes d'ACI diminuent l'information mutuelle (donc la redondance d'information) entre les composantes d'un vecteur.

Nous proposons d'améliorer les codeurs d'images satellitales multi-composantes, embarqués et au sol, par l'étude de transformations inversibles ou réversibles, basées sur des décompositions multi-résolutions et l'ACI. Un des objectifs est d'évaluer l'utilité d'un codage progressif multirésolution allant jusqu'au sans perte, en regard des contraintes que cette propriété pourrait introduire. Globalement, les problèmes à résoudre sont divers : étendre l'ACI à des données discrètes ; construire une "bonne" transformation réversible d'images multi-composantes ; étendre le codage par "arbres de zéros" aux images multispectrales ; construire un "bon" codeur entropique. Pour atteindre l'objectif, nous réunissons différentes équipes de recherche dont les compétences sont complémentaires et recouvrent tous les domaines concernés. »

Ce projet a permis le financement d'une allocation de recherche pour la thèse en mathématiques appliquées d'I. P. Akam Bita, intitulée *Sur une approche de l'analyse en composantes indépendantes à la compression des images multicomposantes*, soutenue en février 2007 à l'Université Joseph Fourier de Grenoble, et dont j'étais co-directeur (50% d'encadrement), avec D. T. Pham (50% d'encadrement). Elle a donné lieu aux publications suivantes :

- deux articles, un long et un court, dans la revue interbationale *Signal Processing* [4, 5] ;
- quatre congrès internationaux avec actes et comité de lecture [17, 20, 21, 23] ;
- trois ateliers avec actes sans comité de lecture [37, 38, 39].

Avec I. P. Akam Bita et D.-T. Pham, nos travaux de recherche ont porté dans un premier temps sur les expressions du gain de codage généralisé quand les transformations sont appliquées après (ou avant) des décompositions en ondelettes bi-dimensionnelles, et sur la recherche d'images multispectrales ou hyperspectrales satellitaires pour lesquelles les gains de codage des transformations retournées par les algorithmes `OrthICA` et `GCGsup` sont sensiblement supérieurs à celui de la transformation de Karhunen-Loève. C'est par exemple le cas d'images hyperspectrales AVIRIS, comme cela a été présenté à une conférence internationale avec comité de lecture [20], à un atelier organisé par le CNES [38] et aux journées PaRI-STIC 2005 [39]. Il s'avère que les algorithmes `OrthICA` et `GCGsup` doivent être modifiés pour donner la transformation (orthogonale ou non) qui maximise le gain de codage généralisé associé à un schéma de compression compatible avec le standard JPEG2000 Partie 2. Ces résultats ont fait l'objet d'une publication dans un congrès international avec comité de lecture [17] et d'un papier long [5] accepté pour publication dans la revue *Signal Processing*. Puis nos travaux se sont poursuivis par la recherche de l'expression du gain de codage généralisé associé à un mélange convolutif (c'est-à-dire une opération de filtrage), étendant ainsi l'expression du gain de codage généralisé associé à un mélange instantané (autrement dit une simple transformation linéaire), et à la mise au point d'un algorithme de maximisation de ce gain de codage par une méthode quasi Newton de descente de gradient dite BFGS. Ces résultats ont fait l'objet d'une présentation dans un *workshop* [21], d'un poster aux journées Pari-STIC 2006 [37] et d'un papier long en préparation à soumettre pour publication dans une revue internationale [49].

Les codecs obtenus avec les transformations spectrales optimales font mieux que l'état de l'art [5], mais au prix d'une plus grande complexité. Nous avons alors cherché à réduire la complexité du calcul

de la transformation optimale. Une première approche, consiste à modifier le critère d'optimalité en supposant les données gaussiennes (donc seuls les moments d'odres 1 et 2 sont à estimer), tout en gardant le schéma de compression compatible avec le standard JPEG2000 Partie 2, où la redondance spatiale est réduite par des décompositions en ondelette 2D et la redondance spectrale au moyen d'une transformation linéaire. Un algorithme de diagonalisation jointe amélioré par D. T. Pham (`JADO`) permet alors de retourner une transformation spectrale orthogonale qui minimise ce critère. Cette étude a donné lieu à la publication d'un article court dans la revue *Signal Processing* [4].

Suite au projet ACI^{2}M, le CNES a financé le post-doc de Mohamed Hariti d'une durée de deux ans (2007–2008), pour étudier un codeur par arbres de zéros 3-D bien adapté aux transformations à base d'ACI. J'ai co-encadré (35% d'encadrement) ce post-doc avec Carole Thiebaut (10% d'encadrement) et Emmanuel Christophe (10% d'encadrement) du CNES, Pierre Duhamel (10% d'encadrement) du LSS, et Jean-Louis Gutzwiller (35% d'encadrement) de Supélec. Les résultats obtenus ont fait l'objet d'un rapport de recherche [43], de deux présentations orales, l'une à un séminaire du CNES [36] et l'autre à une conférence avec actes et comité de lecture [15], et d'un papier long [50] soumis pour publication dans une revue.

Toujours en continuité du projet ACI^{2}M, et grâce à Michel Narozny qui a monté et proposé avec la société LUXspace un dossier accepté par l'ESA (*European Space Agency*) pour financer une *Innovation Triangle Initiative* (ITI) de type *Demonstration of Feasibility & Use*, nous avons été partenaires de cette ITI, c'est le projet HyperComp, dont le but était de développer et de prouver la faisabilité et l'utilisation — dans un environnement de laboratoire — d'un démonstrateur pour un système de compression d'images multi-composantes satellitaires embarqué utilisant les transformations optimales à base d'ACI. Le projet HyperComp réunissait trois partenaires :

- la société luxembourgeoise LUXspace, avec MM Jochen Harms, Florio Dalla Vedova et Isidore Paul Akam Bita ;
- la société allemande OHB System AG ;
- Supélec avec Jean-Louis Gutzwiller et moi-même.

Pour qu'un codeur basé sur les transformations spectrales optimales puisse être embarqué à bord d'un satellite, nous devions réduire significativement la complexité du calcul de la transformation spectrale. Nous avons adopté une autre approche que celle mentionnée ci-dessus et qui consiste à calculer une fois pour toutes une transformation spectrale quasi-optimale sur une base d'apprentissage constituée d'images multicomposantes issues d'un seul capteur et à l'appliquer sur d'autres images provenant du même capteur. Ce projet HyperComp a donné lieu à trois présentations dans des conférences internationales avec actes et comité de lecture [13, 14, 16] et à la préparation d'un article [51] soumis pour publication dans la revue électronique *Journal of Applied Remote Sensing*. De plus, compte-tenu des bons résultats obtenus, une demande de co-financement par l'ESA de la phase 3 (i.e., de l'implantation d'un prototype de codeur embarqué) de l'ITI HyperComp a été envoyée par LUXspace. Cette étape ne correspondant plus à de la recherche, même appliquée, Supélec n'est plus partenaire, mais nous pouvons avoir la satisfaction de voir que, sous réserve de l'acceptation par l'ESA de la proposition, nos recherches vont se matérialiser dans un prototype de système de compression embarqué à bord d'un satellite. Un article regroupant les résultats obtenus avec le post-doc CNES et le projet HyperComp est en préparation pour être soumis dans une revue internationale avec comité de lecture.

Mes charges d'enseignement à Supélec portent essentiellement sur le traitement statistique du signal en 2ème et 3ème année (voir CV joint), le cours de 3ème année est commun à Supélec, au Master Recherche de mathématiques fondamentales et appliquées cohabilité par Supélec et l'Université Paul Verlaine de Metz et au *Master of Science* de Georgia Tech Lorraine. Pendant la vingtaine d'années que j'ai passée comme enseignant-chercheur à Supélec, j'ai participé à beaucoup d'enseignements

(travaux dirigés, cours ou travaux de laboratoires) portant sur les mathématiques appliquées, j'ai rédigé un livre [55], quelques polycopiés ([62, 59, 61, 60, 58]), quelques dossiers pour l'épreuve de TIPE (Travaux d'Initiative Personnelle Encadrés) aux concours d'entrées aux Grandes Écoles (après avoir été examinateur, je suis responsable pédagogique en mathématiques et informatique depuis octobre 2006) et de nombreux sujets d'examens, de travaux de laboratoires et de travaux dirigés. J'ai également encadré ou co-encadré des élèves de troisième année en conventions d'études industrielles (CEI) sur les sujets suivants :

- "Étude de la séparation des signaux en sismique simultanée" pour Total avec Mle Leclercq (élève) et MM. J.-L. Collette et J.-L. Boelle pour le co-encadrement, 2010.
- "Étude et mise au point d'un procédé de calibration pour un imageur matriciel", pour Sagem Defense Sécurité avec M. Derouvroy (élève) et M. Gardette pour le co-encadrement, 2010.
- "Analyse statistique du bruit de détecteurs matriciels dédiés à l'imagerie infrarouge", pour Sagem Défense Sécurité avec M^{lle} Tondriaux et M. de Torres Garcia (élèves), et MM. Gardette et Genty pour le co-encadrement, 2009.
- "Détection des sursauts gamma", pour le CEA (centre de Saclay) avec MM. Courtois et Seurrat de la Boulaye (élèves) et MM. Schanne et Pietquin pour le co-encadrement, 2008.
- "Étude des dispersions des données sur une chaîne de production d'acier Dual Phase", pour la société ARCELOR MITTAL-RESEARCH SA, avec MM. Maugey et Abbas-Turki (élèves), et MM. Fricout et Pietquin pour le co-encadrement, 2007.
- "Simulateurs de jeux basés sur les prévisions de marchés boursiers" pour la société LUSIS SA avec MM Badouin (élève) et Popineau (co-encadrant), 2006 ;
- "Augmentation de la bande passante d'un moyen de mesure" pour la société MICHELIN avec MM. Froger et Lemercier (élèves) et Collette (co-encadrant), 2006 ;
- "Traitement des signaux de fluctuations de pression issus de simulations numériques" pour la société RENAULT avec MM. Apostolou et Mauger (élèves) et MM. Morosini et Illy (co-encadrants), 2006 ;
- "Traitement des signaux de fluctuations de pression en paroi générées par un écoulement turbulent" pour la société RENAULT avec Mme Bouquet, MM. Bitton et Riahi (élèves) et M^{lle} Arguillat, MM. Ricot et Morosini (co-encadrants), 2005 ;
- "Recherche de représentations optimales pour l'interception de signaux radar" pour la société THALES Systèmes Aéroportés avec MM. Bect, Bloch, Bondier et Huygen (élèves) et M^{lle} Chevaillier et M Delabbaye (co-encadrants), 2003 ;
- "Validation d'une méthode d'identification par analyse en composantes principales" pour la société MICHELIN, avec Mlle Neyme et MM. Artault, Bailloeul, Mérit et Moratal y Perez (élèves) et MM. Biesse et Collette (co-encadrants), 2001.

J'ai réalisé des études industrielles :

- "Analyse statistique de données pour les applications *Galvallia* et *Skin pass*" pour la société ARCELOR (en collaboration avec C. Mailhan, J.-L. Collette et M. Ianotto) 2004 ;
- étude confidentielle pour un équipementier automobile, 2002 ;
- "Estimation des intervalles de confiance dans un perceptron multi-couches" pour la société IRSID (en collaboration avec M. Ianotto) 2002 ;

et j'ai co-encadré une dizaine de stagiaires du Master Recherche (ou DEA) de mathématiques appliquées.

Chapitre 2

Sur la stabilité des filtres récursifs bi-dimensionnels

2.1 Introduction

Mon travail de thèse, sous la direction de M. Benidir, a porté sur le test de la stabilité des filtres numériques récursifs bi-dimensionnels avec l'objectif d'apporter une solution pratique qui soit plus satisfaisante que celles existantes. Pour cela, nous avons naturellement découpé mes investigations en deux parties. La première, dont les résultats sont exposés au chapitre II du mémoire de thèse [48] et résumés à la section suivante, a consisté en la recherche d'un critère, c'est-à-dire une condition nécessaire et suffisante, de stabilité, et la deuxième, objet du chapitre III de [48] et résumée deux sections plus loin, a porté sur l'algorithme proprement dit qui est une traduction du critère précédent en langage informatique.

Plus précisément, un nouveau critère de stabilité a été établi, puis comparé à ceux existants. Après avoir proposé une définition objective et précise de la notion intuitive de complexité d'un tel critère, il a été montré que celui introduit dans la thèse a la même que d'autres classiques (Huang-Goodman, Strintzis) et qu'il est impossible de trouver un critère de complexité moindre. Ces résultats reposent sur l'étude, dans l'espace des polynômes complexes à une variable et de degré limité, du domaine de stabilité 1-D, qui consiste en l'ensemble des polynômes apparaissant aux dénominateurs des fonctions de transfert des filtres dynamiques causaux et stables. L'équation—dont l'irréductibilité est démontrée—de la plus petite hypersurface contenant la frontière de ce domaine est donnée [10]. Il est également observé que cette hypersurface ne pénètre pas à l'intérieur du domaine. Ces résultats apportent un tout autre éclairage au problème de la stabilité des filtres numériques récursifs bi- ou multi-dimensionnels et redonnent en particulier les critères classiques.

Un autre résultat de la thèse, établi au chapitre II de [48], est que tout algorithme, décidant en un nombre fini de pas si un filtre numérique récursif causal bi-dimensionnel quelconque (qu'il soit réel ou complexe) est stable—entrée bornée-sortie bornée—, requiert le test de nullité d'un résultant particulier (c'est inévitable). L'examen des algorithmes existants montre qu'ils commencent tous par calculer cette expression—ou une autre équivalente—avant de tester si elle s'annule.

Il est également proposé dans [48] deux procédures permettant de décider, en un nombre fini de pas, si un filtre numérique récursif tri-dimensionnel est stable ou non et si tous les polynômes d'une famille convexe, engendrée par un nombre fini de polynômes à une indéterminée, ont tous leurs zéros à l'intérieur strict du cercle unité [31].

Le chapitre III du rapport de thèse est consacré à la présentation d'un algorithme de test de stabilité pour les filtres numériques récursifs bi-dimensionnels. Le nombre d'opérations requises par

cet algorithme "rapide" diffère d'un—voire plusieurs—ordre de grandeur par rapport aux autres algorithmes connus en 1993. Les filtres traités sont supposés réels, dépourvus de toute singularité non essentielle de la deuxième espèce sur le bi-cercle unité, quart-plans causals ou demi-plans non symétriques semi-causals. L'algorithme est basé sur la condition nécessaire et suffisante mentionnée ci-dessus. Cette dernière fait intervenir le calcul du résultant de deux équations à deux indéterminées, obtenu en éliminant entre elles une des deux inconnues. Le résultant est une expression "classique", il est défini et étudié dans la théorie de l'élimination et dans les cours élémentaires d'algèbre. Il s'exprime sous différentes formes de déterminants (résultant de Sylvester, résultant de Bézout). Comme résultat de la thèse, il a été établi, que dans le cas qui intéresse l'étude de la stabilité, le résultant à calculer est égal au déterminant d'une matrice proche de Toeplitz, dont le rang de déplacement vaut 2. Cela a autorisé l'usage, pour la première fois à notre connaissance dans un test de stabilité [11], d'un algorithme rapide d'inversion de matrices, celui de Levinson-Szegö généralisé.

L'étude de la stabilité des filtres numériques récursifs bi- ou multi-dimensionnels est grandement facilitée par une bonne connaissance des résultats relatifs à la stabilité de tels filtres mono-dimensionnels, non seulement parce que les chercheurs en ce domaine ont souvent étudié au préalable le cas 1-D et que leurs articles y font référence, mais surtout parce que ces problèmes sont intrinsèquement liés, comme le montre, par exemple, l'éclairage du chapitre II de [48] (voir également [2] ou [1]). La littérature sur le sujet est très volumineuse. Elle s'étale dans le temps depuis environ 1830 (Cauchy-Sturm) jusqu'à aujourd'hui et recoupe différentes disciplines des sciences de l'ingénieur (automatique, traitement du signal) et des mathématiques pures (algèbre, analyse, géométrie algébrique) et appliquées (équations différentielles, valeurs propres, informatique). Un gros travail de synthèse a été réalisé [47] avec M. Benidir pour rédiger le livre [1].

Il est d'usage en traitement du signal ou en automatique de proposer des programmes informatiques dont les calculs sont faits avec une arithmétique à virgule flottante et l'algorithme présenté au chapitre III de [48] (voir également [11] ou [3]) est de ce type. L'une de ses étapes consiste à vérifier si un polynôme réel de degré important (≥ 20) s'annule sur le segment réel $[-2; 2]$. Il est d'ailleurs montré, que cette vérification est inévitable dans tout test de stabilité de filtres récursifs bi-dimensionnels, soit sous cette forme, soit sous une autre équivalente. Il est bien connu que la localisation des zéros d'un polynôme peut être très sensible à de minimes fluctuations de ses coefficients, ceci d'autant plus que son degré est élevé. Cela a pour conséquence, du fait des erreurs d'arrondi inévitables en dehors du calcul formel, que l'algorithme établi au chapitre III de [48], comme tous ceux équivalents proposés en traitement du signal, n'est pas fiable à cent pour cent. La question suivante se pose alors naturellement : *étant donné un polynôme développé en série entière, est-il possible de mesurer à quel point la localisation de ses zéros est sensible à de petites fluctuations de ses coefficients ?* Ce problème est un cas particulier de celui de la localisation des valeurs propres d'une matrice dont les coefficients sont soumis à de faibles perturbations : il suffit d'associer au polynôme sa matrice compagnon. Ce fut l'objet d'importantes recherches en mathématiques pures et appliquées depuis le début du siècle et il existe d'excellents ouvrages de synthèse allant jusqu'aux années 60 (Wilkinson, Parodi). Après la thèse, nous avons comparé différents algorithmes de test de stabilité de filtres 2-D, face aux erreurs d'arrondis [9] (voir également [1, 3] et une section ci-après), il apparaît que la quasi totalité des algorithmes sont très sensibles aux erreurs d'arrondi dès que le plus petit des degrés du polynôme est supérieur strictement à deux, sauf un que nous avons proposé dans [3] (voir aussi [1] pp. 231–235) et qui ne marche que sous certaines conditions peu restrictives en pratique (voir l'avant dernière section de ce chapitre).

2.2 Optimalité des critères de stabilité 2-D

Cette section présente les résultats obtenus pendant ma thèse, afin de faciliter la lecture des sections suivantes qui portent, elles, sur des résultats obtenus après ma thèse. Cette section commence par l'étude du domaine de stabilité des filtres numériques récursifs mono-dimensionnels dans l'espace des polynômes apparaissant au dénominateur de leur fonction de transfert. Il est établi que pour des filtres à coefficients complexes, ce domaine est connexe, et l'équation de la plus petite hypersurface contenant sa frontière est donnée. Il est également montré que cette hypersurface ne pénètre pas à l'intérieur du domaine de stabilité.

Les propriétés géométriques du domaine de stabilité—dans l'espace affine des polynômes complexes de degré n'excédant pas n—que nous avons révélées [10], offrent un nouvel éclairage au problème de la stabilité des filtres numériques récursifs bi- ou multi-dimensionnels. Elles permettent en particulier de retrouver les critères classiques (nous nous focaliserons sur le cas 2-D) et nous ont permis d'en construire un nouveau en 1993.

Après avoir précisé la notion intuitive de complexité d'un tel critère, il est montré dans la thèse que celui que nous avions introduit a une complexité minimale : aucun critère de test de stabilité 2-D ne peut avoir une complexité moindre.

2.2.1 Domaine des polynômes 1-D stables

Frontière du domaine de stabilité

Considérons un polynôme complexe général, i. e., dont les coefficients sont des indéterminées,

$$P = a_0 x^n + a_1 x^{n-1} + \cdots + a_n \tag{2.1}$$

de degré au plus n. Dans la suite, le coefficient a_0 pourra s'annuler. Soit P^* le polynôme déduit de P par la relation

$$P^* = \overline{a}_n x^n + \overline{a}_{n-1} x^{n-1} + \cdots + \overline{a}_0, \tag{2.2}$$

où $\overline{a}_k$ désigne le conjugué du nombre complexe a_k. Si P est de degré n ($a_0 \neq 0$), P^* est le *polynôme réciproque* de P. Par convention, quand le degré k de P est inférieur à n, nous disons que l'infini (∞) est un zéro de P avec un ordre de multiplicité (odm) de $n-k$. Avec cette convention, tout polynôme P, défini par la relation (2.1), non identiquement nul, admet exactement n zéros comptés suivant leurs odm, ∞ compris.

Au polynôme P ci-dessus, associons le point de $\mathbb{C}^{n+1}$, encore noté P, admettant $(a_0, a_1, \ldots, a_n)$ comme coordonnées dans le repère canonique. Au moyen de cette bijection, l'espace vectoriel $\mathbb{C}_n[x]$ des polynômes complexes de degré au plus n est identifié à l'espace affine $\mathbb{C}^{n+1}$. Pour tout entier p ($0 \leq p \leq n$), construisons le sous-ensemble $\mathcal{D}_p$ de $\mathbb{C}^{n+1}$ constitué de tous les polynômes ayant exactement p zéros à l'intérieur ou sur le cercle unité (CU) et $n-p$ (dont éventuellement ceux situés en ∞) à l'extérieur strict de ce cercle. Les sous-ensembles $\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_n$ forment une partition de $\mathbb{C}^{n+1} \setminus \{0\}$. Les premières propositions donnent quelques propriétés topologiques et géométriques de ces sous-ensembles, nous omettons les démonstations qui peuvent être trouvées dans le rapport de thèse ou dans [1].

Proposition 1 *Les sous-ensembles $\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_n$ sont connexes.*

Introduisons encore quelques définitions et notations. Soit

$$a_k = \alpha_k + i\beta_k \qquad (0 \leq k \leq n)$$

la décomposition suivant ses parties, réelle ou imaginaire, du k-ème coefficient du polynôme (2.1). Identifions le plan complexe $\mathbb{C}$ avec l'espace affine réel $\mathbb{R}^2$, et l'espace $\mathbb{C}^{n+1}$ des coefficients $(a_0, \ldots, a_n)$ du polynôme (2.1) avec l'espace affine réel $\mathbb{R}^{2n+2}$ de leurs parties réelles et imaginaires $(\alpha_0, \ldots, \alpha_n, \beta_0, \ldots, \beta_n)$. Si $\mathbf{F}$ est un polynôme réel aux $2n+2$ variables $\alpha_0, \ldots, \alpha_n, \beta_0, \ldots, \beta_n$, et si $\mathbf{F}$ n'est pas constant, alors le sous-ensemble de $\mathbb{R}^{2n+2}$ contenant tous les zéros **réels** de $\mathbf{F}$ est appelé *l'hypersurface d'équation*

$$\mathbf{F}(\alpha_0, \ldots, \alpha_n, \beta_0, \ldots, \beta_n) = 0.$$

Dans la suite, deux types de polynômes seront utilisés : certains complexes et à une seule variable, comme P et P^*, d'autres réels et aux $2n+2$ variables $\alpha_0, \ldots, \alpha_n, \beta_0, \ldots, \beta_n$, comme $\mathbf{F}$. Ces derniers seront toujours désignés par des majuscules en caractères gras.

Pour $0 \leq p \leq n$, soit $\partial\mathcal{D}_p$ la frontière de $\mathcal{D}_p$, constituée de l'ensemble des points de $\mathbb{C}^{n+1}$ appartenant à la fermeture de $\mathcal{D}_p$, sans être en son intérieur. Un résultat de la thèse est d'avoir donné l'équation de la plus petite hypersurface contenant $\partial\mathcal{D}_0$ et d'avoir montré qu'elle ne pénètre pas à l'intérieur de $\mathcal{D}_0$. Pour cela, trois résultats intermédiaires ont été établis. Le premier indique que l'union $\mathcal{B}_1$ des frontières $\partial\mathcal{D}_p$, pour $0 \leq p \leq n$, est égale à l'ensemble des polynômes $P \in \mathbb{C}^{n+1}$ qui s'annulent sur le cercle unité. Le deuxième révèle que cette union $\mathcal{B}_1$ est incluse dans l'hypersurface $\mathcal{B}_2$ d'équation

$$\mathbf{R}(P, P^*) = 0, \tag{2.3}$$

où $\mathbf{R}(P, P^*)$ est le résultant des polynômes P et P^* et que l'hypersurface $\mathcal{B}_2$ ne pénètre ni à l'intérieur de $\mathcal{D}_0$ ni à l'intérieur de $\mathcal{D}_n$, mais qu'elle coupe l'intérieur de tous les autres sous-ensembles $\mathcal{D}_p$ $(0 < p < n)$. Enfin, le troisième montre que l'équation (2.3) est irréductible, c'est-à-dire qu'elle ne peut pas être factorisée, dans le cas général où les coefficients de P sont indéterminés.

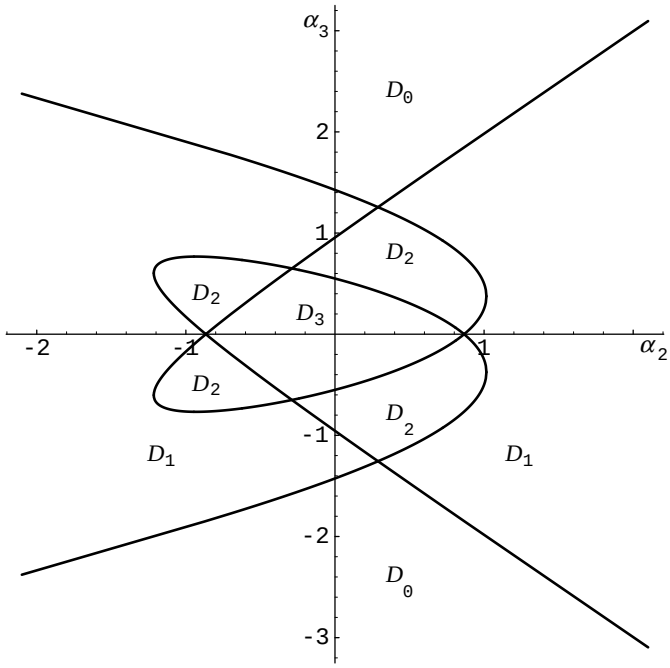


Fig. 1 $\quad P = x^3 + (\alpha_2 + 0.5i)x + \alpha_3$

Exemple 1 Les figures 1 et 2 montrent deux sections planes des sous-ensembles $\mathcal{D}_p$ quand $n = 3$.

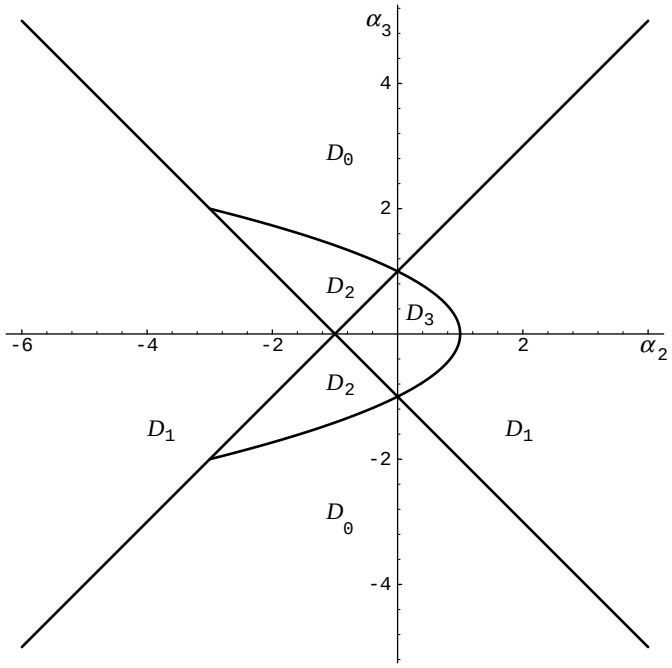


Fig. 2 $\quad P = x^3 + \alpha_2 x + \alpha_3$

Il est bien connu que les zéros d'un polynôme sont des fonctions continues de ses coefficients et inversement (cela reste vrai avec la convention adoptée ci-dessus sur les zéros à l'infini). Il résulte d'une telle continuité, que pour $0 \leq p \leq n$, le sous-ensemble $\mathcal{D}'_p$, formé de tous les polynômes $P \in \mathbb{C}^{n+1}$ admettant exactement p zéros à l'intérieur (strict) du CU et $n-p$ à l'extérieur strict de ce cercle, est ouvert. D'autre part, tout point de $\mathcal{D}_p$ est la limite d'une suite de points appartenant à $\mathcal{D}'_p$, donc $\mathcal{D}'_p$ est l'intérieur de $\mathcal{D}_p$. D'après les définitions, il est évident que $\mathcal{D}_0$ est ouvert, donc $\mathcal{D}'_0 = \mathcal{D}_0$.

Proposition 2 *L'union $\mathcal{B}_1$ de toutes les frontières $\partial\mathcal{D}_p$, pour $0 \leq p \leq n$, est égale à l'ensemble des polynômes $P \in \mathbb{C}^{n+1}$ qui s'annulent sur le cercle unité.*

Un résultat classique d'algèbre affirme que quand $a_0 \neq 0$ ou $\overline{a}_n \neq 0$, les polynômes P et P^* ont un zéro en commun, si et seulement si, leur résultant $\mathbf{R}(P, P^*)$ s'annule. Dans ce cas, le zéro en commun est fini. Avec la convention adoptée pour les zéros à l'infini, l'hypothèse $a_0 \neq 0$ ou $\overline{a}_n \neq 0$ précédente est inutile. L'expression $\mathbf{R}(P, P^*)$ est un polynôme réel des variables $\alpha_0, \ldots, \alpha_n$, $\beta_0, \ldots, \beta_n$ (voir la proposition 5 ci-dessous), ainsi, l'équation (2.3) définit bien une hypersurface $\mathcal{B}_2$. Enfin, si P s'annule sur le cercle unité en ω, alors $P^*(\omega) = 0$ puisque $1/\overline{\omega} = \omega$ et que les zéros de P^* s'obtiennent à partir de ceux de P par passage à l'inverse du conjugué. Donc $\mathcal{B}_2$ contient $\mathcal{B}_1$. Nous venons de trouver une hypersurface, $\mathcal{B}_2$, contenant $\partial\mathcal{D}_0$ ainsi que toutes les autres frontières $\partial\mathcal{D}_p$, $(0 \leq p \leq n)$.

Proposition 3 *Pour $0 \leq p \leq n$, $\mathcal{B}_2 \cap \mathcal{D}'_p = \emptyset$, si et seulement si, $p = 0$ ou $p = n$.*

Par conséquent, $\mathcal{B}_2$ "reste" sur la frontière de $\mathcal{D}_0$ et sur celle de $\mathcal{D}_n$. Nous en déduisons alors immédiatement, que restreints au sous-ensemble fermé $\Delta_0 = \mathcal{D}_0 \cup \partial\mathcal{D}_0$, les trois domaines $\mathcal{B}_2$, $\mathcal{B}_1$ et $\partial\mathcal{D}_0$ coïncident exactement, i. e., $\mathcal{B}_2 \cap \Delta_0 = \mathcal{B}_1 \cap \Delta_0 = \partial\mathcal{D}_0$. De même, si $\Delta_n = \mathcal{D}_n \cup \partial\mathcal{D}_n$, alors $\mathcal{B}_2 \cap \Delta_n = \mathcal{B}_1 \cap \Delta_n = \partial\mathcal{D}_n$. Il en résulte la proposition suivante qui trouve d'intéressantes applications dans l'étude de la stabilité des filtres numériques récursifs 1-D, 2-D et N-D.

Proposition 4 *Partant d'un point P_0 dans $\mathcal{D}'_0$ (resp. $\mathcal{D}'_n$), si P varie continûment*[1]*, alors il intersecte $\partial\mathcal{D}_0$ (resp. $\partial\mathcal{D}_n$), si et seulement s'il coupe l'hypersurface $\mathcal{B}_2$, si et seulement s'il intersecte $\mathcal{B}_1$.*

Preuve. On a $\partial\mathcal{D}_0 \subset \mathcal{B}_1 \subset \mathcal{B}_2$. Considérons une fonction continue f d'un sous-ensemble non vide connexe $I \subset \mathbb{R}^r$, ($r \geq 1$ fixé), dans $\mathbb{C}^{n+1}$, telle que pour $t_0 \in I$ fixé, $f(t_0) = P_0 \in \mathcal{D}'_0$. Supposons qu'il existe $t_1 \in I$ vérifiant $f(t_1) \in \mathcal{B}_2$. I étant connexe, il existe une fonction continue g de $[0;1]$ dans I, telle que $g(0) = t_0$ et $g(1) = t_1$. L'ensemble

$$\{\mu \in [0;1] : \forall x \in [0;\mu], f \circ g(x) \in \mathcal{D}'_0\}$$

est non vide (il contient 0) et majoré par 1. Il admet donc une borne supérieure $\tau \leq 1$, qui vérifie $f \circ g(\tau) \in \partial\mathcal{D}_0$. Le même raisonnement s'applique à $\mathcal{D}_n$.

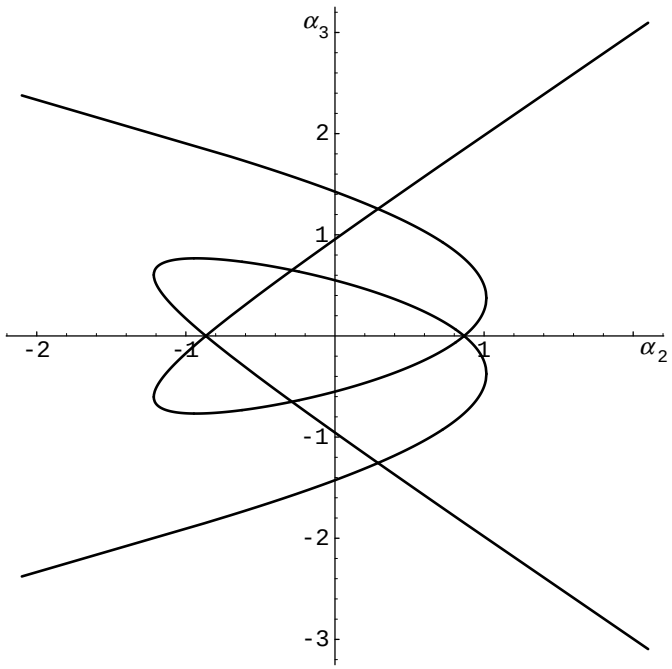


Fig. 3 $P = x^3 + (\alpha_2 + 0.5i)x + \alpha_3$

Exemple 2 Les figures 3 et 4 montrent deux sections planes de $\mathcal{B}_2$ quand $n = 3$. En comparant entre-elles les figures de même parité : 1 avec 3 et 2 avec 4, remarquons que la parabole de Fig. 4 pénètre à l'intérieur de $\mathcal{D}'_1$. Toutefois, $\mathcal{B}_2$ ne peut couper ni $\mathcal{D}'_0$, ni $\mathcal{D}'_n$.

[1] c'est à dire s'il existe un sous-ensemble non vide connexe $I \subset \mathbb{R}^r$ ($r \geq 1$ fixé), $t_0 \in I$, et une application continue f de I dans $\mathbb{C}^{n+1}$, tels que pour $t \in I, P = f(t)$ et $P_0 = f(t_0)$.

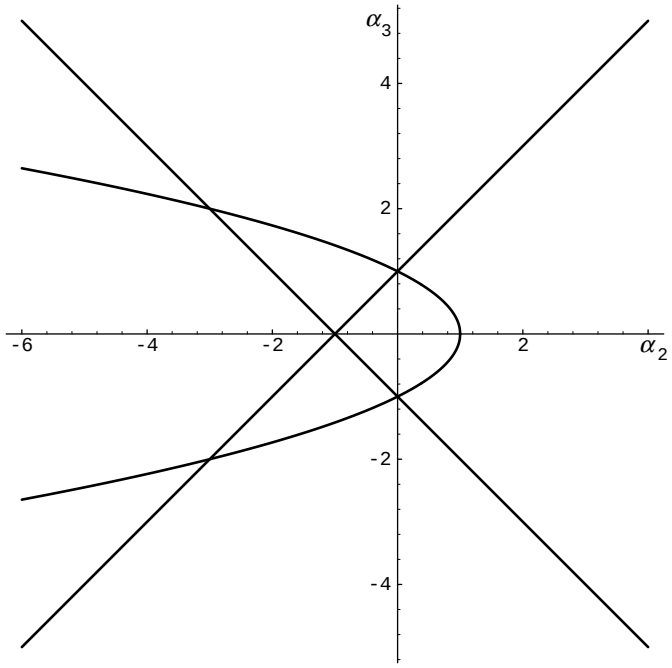


Fig. 4 $P = x^3 + \alpha_2 x + \alpha_3$

Pour terminer l'étude du domaine de stabilité 1-D, énonçons l'irréductibilité de l'équation (2.3), dans le cas général où les coefficients du polynôme P, défini par la relation (2.1), sont indéterminés. Il est bien connu que le résultant de Sylvester de deux polynômes généraux indépendants, f et g, est un polynôme irréductible dont les variables sont les coefficients de f et de g. Dans le cas qui nous intéresse, les coefficients de P^* pouvant s'exprimer en fonction de ceux de P, il n'est pas évident que $\mathbf{R}(P, P^*)$, polynôme en $\alpha_0, \ldots, \alpha_n$, $\beta_0, \ldots, \beta_n$, soit irréductible, c'est cependant le cas.

Proposition 5 *Quand les coefficients $a_k = \alpha_k + i\beta_k$ $(0 \le k \le n)$ du polynôme P défini en (2.1) sont des indéterminées complexes décomposées suivant leurs parties, réelle ou imaginaire, le résultant $\mathbf{R}(P, P^*)$ de P et P^*, défini en (2.2), est un polynôme réel irréductible, aux variables $\alpha_0, \ldots, \alpha_n$, $\beta_0, \ldots, \beta_n$.*

Exemple 3 Quand $n = 2$,

$$\begin{aligned}\mathbf{R}(P, P^*) &= |a_0|^4 - |a_0|^2|a_1|^2 + a_0 a_2 \overline{a}_1^2 + \\ &\quad \overline{a}_0 \overline{a}_2 a_1^2 - 2|a_0|^2|a_2|^2 - |a_1|^2|a_2|^2 + |a_2|^4.\end{aligned}$$

Cette expression visiblement réelle, se développe en un polynôme à coefficients entiers en $\alpha_0, \ldots, \alpha_n$, $\beta_0, \ldots, \beta_n$ dont l'écriture demanderait plusieurs lignes.

Quand le polynôme P est réel, le résultant de P et de P^* n'est plus irréductible, comme l'avait remarqué Cohn et d'autres avant lui. La proposition précédente n'est alors plus valide. Ce résultat peut se retrouver par transformation homographique : l'équation

$$(1+x)^n P\Big((1-x)/(1+x)\Big) = f(x^2) + x g(x^2)$$

définit deux polynômes réels f et g qui vérifient

$$(-1)^{n(n+1)/2} P(1) P(-1) \Big(\mathbf{R}(f, g)\Big)^2 = 2^{n(n-1)} \mathbf{R}(P, P^*).$$

Cette équation redonne un résultat de Liénard et Chipart.

Proposition 6 *Pour tout entier p $(0 \leq p \leq n)$, $\mathcal{B}_2$ est la plus petite hypersurface contenant $\partial\mathcal{D}_p$.*

Cette dernière proposition peut s'exprimer ainsi : il n'existe pas de famille finie d'équations polynomiales, aux indéterminées $\alpha_0, \ldots, \alpha_n, \beta_0, \ldots, \beta_n$ et ayant chacune un degré inférieur à celui de l'équation (2.3), qui définisse une hypersurface contenant $\partial\mathcal{D}_p$.

Applications aux critères de stabilité

Nous supposons dans ce paragraphe que chaque coefficient a_k du polynôme P, défini en (2.1), est une fonction continue de variables réelles $\underline{v} = (v_1, \ldots, v_r) \in I$, où $I = I_1 \times I_2 \times \cdots \times I_r$ est le produit cartésien d'intervalles réels non vides. Notons $(P_v)_{v\in I}$ la famille de tous les polynômes obtenus avec ces fonctions continues quand $v \in I$. Nous dirons d'une telle famille de polynômes qu'elle est *continue*. Puisque I est un sous-ensemble connexe de $\mathbb{R}^r$, la famille continue de polynômes $(P_v)_{v\in I}$ forme un sous-ensemble connexe de $\mathbb{C}^{n+1}$.

Proposition 7 *Soit $(P_v)_{v\in I}$ une famille continue de polynômes complexes de degré n'excédant pas n. Soit, pour $v \in I$, P_v^* défini par la relation (2.2) quand P_v est écrit sous la forme (2.1).*

(i) Tout polynôme P_v de la famille est dans $\mathcal{D}_0$, si et seulement s'il existe $v_0 \in I$ tel que $P_{v_0} \in \mathcal{D}_0$ et le résultant $\mathbf{R}(P_v, P_v^*)$ de P_v et P_v^* ne s'annule pas pour tout $v \in I$.

(ii) S'il existe $v_0 \in I$ tel que $P_{v_0} \in \mathcal{D}_0$ et si le résultant $\mathbf{R}(P_v, P_v^*)$ de P_v et P_v^* s'annule pour $v \in I$, alors il existe $v_1 \in I$ tel que le polynôme P_{v_1} s'annule sur le cercle unité.

Ces résultats restent valides si $\mathcal{D}_0$ est remplacé par l'intérieur $\mathcal{D}'_n$ de $\mathcal{D}_n$.

Preuve. Cela résulte directement de la proposition 4.

L'assertion (i) de la proposition précédente peut être interprétée en termes d'algorithmes. Fixons à la fois $r \geq 1$, le pavé $I \subset \mathbb{R}^r$ d'intérieur non vide, $v_0 \in I$ et $n > 0$ un entier naturel. Notons $\mathcal{P}$ l'ensemble de toutes les familles continues $(P_v)_{v\in I}$ de polynômes complexes dont le degré n'excède pas n et $\mathcal{P}_1$ le sous-ensemble de $\mathcal{P}$ des familles $(P_v)_{v\in I}$ pour lesquelles $P_{v_0} \in \mathcal{D}_0$. Nous considérons trois algorithmes $\mathcal{A}$, $\mathcal{B}$ et $\mathcal{C}$, finissant chacun en un nombre *fini* de pas[2]. Les deux premiers admettent comme entrée tout élément de $\mathcal{P}$, alors que toutes les entrées admises par le dernier forment le sous-ensemble $\mathcal{P}_1$. Ces trois programmes répondent par oui ou par non aux questions respectives suivantes, où $(P_v)_{v\in I}$ est une de leurs entrées admissibles quelconque.

$\mathcal{A}$ **:** *est-ce-que tous les éléments de la famille $(P_v)_{v\in I}$ sont dans $\mathcal{D}_0$?*

$\mathcal{B}$ **:** *est-ce-que P_{v_0} appartient à $\mathcal{D}_0$?*

$\mathcal{C}$ **:** *est-ce-que le résultant $\mathbf{R}(P_v, P_v^*)$ s'annule pour $v \in I$?*

Pour ces algorithmes, leur *complexité* est, par définition, égale au nombre d'opérations qu'ils effectuent pour traiter une donnée arbitraire. C'est une fonction de n. Par abus de langage, nous appelons encore *complexité* l'ordre de grandeur de cette fonction quand n tend vers l'infini.

Il existe des algorithmes $\mathcal{B}$ déduits de la règle de Cohn ou de ses dérivées (Marden[3], Jury[4] , ...) ou encore du critère de Routh par transformation homographique et il est clair que tout programme $\mathcal{A}$ ne peut pas avoir une complexité inférieure à celle de $\mathcal{B}$. Il résulte de la proposition précédente,

[2]Nous excluons donc tout algorithme reposant sur un échantillonnage du pavé I.

[3] *The geometry of polynomials*, 2ème édition, Providence, American Mathematical Society, 1966.

[4]"A simplified stability criterion for linear discret systems", *Proceedings of the IRE*, 1493–1500, juin 1962.

que disposant d'un algorithme $\mathcal{B}$, tout algorithme $\mathcal{A}$ devient un algorithme $\mathcal{C}$ et inversement, deux algorithmes $\mathcal{B}$ et $\mathcal{C}$ quelconques forment ensemble un algorithme $\mathcal{A}$. Autrement dit,

$$(\mathcal{A} \text{ et } \mathcal{B}) \Leftrightarrow (\mathcal{C} \text{ et } \mathcal{B})\,.$$

Les algorithmes $\mathcal{A}$ et $\mathcal{C}$ les plus performants ont donc la même complexité et, à l'algorithme $\mathcal{B}$ près, de complexité négligeable, les algorithmes $\mathcal{A}$ et $\mathcal{C}$ sont équivalents.

Intéressons-nous à l'algorithme $\mathcal{C}$. D'après la proposition 5, les coefficients des polynômes P_v pouvant *a priori* prendre n'importe quelles valeurs complexes quand $v \in I$, le résultant $\mathbf{R}(P_v, P_v^*)$ ne peut pas être factorisé *a priori*.

2.2.2 Application aux tests de stabilité des filtres numériques récursifs 2-D

L'application des résultats précédents au problème de la stabilité des filtres numériques récursifs quart-plan causaux, dépourvus de toute singularité non essentielle de la deuxième espèce, permet de retrouver les critères de Huang, Goodman, Shanks, Strintzis, De Carlo, Anderson, Jury ... (voir les livres de Huang[5] ou Dudgeon, Russel et Mersereau[6]).

Soit $P(z_1, z_2)$ un polynôme général complexe à deux variables

$$P(z_1, z_2) = \sum_{h=0}^{n} \sum_{k=0}^{m} a_{h,k} z_1^{n-h} z_2^{m-k} = \sum_{k=0}^{m} a_k(z_1) z_2^{m-k} \quad (m \leq n). \tag{2.4}$$

Il résulte de la proposition 7 (i), avec

$$r = 2, I = [0;1] \times [0;1], v = (\rho, \theta) \in I \text{ et } P_v(z_2) = P(\rho e^{2i\pi\theta}, z_2),$$

que $P(z_1, z_2) \neq 0$ sur le bi-disque unité fermé $|z_1| \leq 1$ et $|z_2| \leq 1$, si et seulement si, $P(0, z_2) \in \mathcal{D}_0$ et le résultant $\mathbf{R}(z_1)$ obtenu par élimination de z_2 entre les deux équations :

$$P(z_1, z_2) = 0 \qquad \text{et} \qquad P^*(z_1, z_2) = \sum_{k=0}^{m} \overline{a_k(z_1)} z_2^k = 0 \tag{2.5}$$

ne s'annule pas sur le dique unité fermé $|z_1| \leq 1$. Il découle également de la proposition 7 (ii) (avec $P_v(z_1) = P(z_1, \rho e^{2i\pi\theta})$ et les mêmes r, v et I que ci-dessus) que $P(z_1, z_2) \neq 0$ sur le bi-disque unité fermé, si et seulement si, les deux conditions suivantes sont satisfaites (Huang-Goodman) :

$$P(z_1, 0) \neq 0 \text{ pour tout } z_1 \text{ tel que } |z_1| \leq 1 \tag{2.6}$$

$$P(z_1, z_2) \neq 0 \text{ pour tout } z_1, z_2 \text{ tels que } |z_1| = 1 \text{ et } |z_2| \leq 1. \tag{2.7}$$

De la même façon, avec $r = 1$, $I = [0;1]$, et $P_v(z_2) = P(e^{2i\pi v}, z_2)$, la condition (2.7) est valide, si et seulement si, les deux conditions suivantes sont vérifiées (Strintzis) :

$$P(1, z_2) \neq 0 \text{ pour tout } z_2 \text{ tel que } |z_2| \leq 1 \tag{2.8}$$

$$P(z_1, z_2) \neq 0 \text{ pour tout } z_1, z_2 \text{ tels que } |z_1| = |z_2| = 1. \tag{2.9}$$

[5] *Two-dimensional digital signal processing I, Topics in applied physics*, **42**, Springer-Verlag, New-York, 1981.
[6] *Multidimensional digital signal processing*, Prentice Hall, 1984.

Dans la relation (2.6) (respectivement (2.8)), le chiffre 0 de "$P(z_1, 0)$" (respectivement le chiffre 1 de "$P(1, z_2)$" peut être remplacé par tout autre point du disque unité fermé (respectivement du cercle unité $|z_1| = 1$).

Nous pouvons facilement étendre ces justifications aux filtres multi-dimensionnels de toute dimension N, pour retrouver le théorème de Strintzis, De Carlo et *al.*

La proposition suivante résulte de la proposition 7 (i).

Proposition 8 *Soit $P(z_1, z_2)$ un polynôme complexe défini par la relation (2.4). La condition (2.7) est valide, si et seulement si, la condition (2.8) est satisfaite et si le résultant $\mathbf{R}(z_1)$, obtenu en éliminant z_2 entre les deux équations (2.5), ne s'annule pas sur le cercle unité.*

Nous en avons déduit un critère de stabilité nouveau en 1993, c'est un résultat de ma thèse.

Proposition 9 (critère de stabilité 2-D) *Soient $P(z_1, z_2)$ un polynôme complexe défini par la relation (2.4) et $\mathbf{R}(z_1)$ le résultant obtenu en éliminant z_2 entre les deux équations (2.5). $P(z_1, z_2)$ ne s'annule pas sur le bi-disque unité fermé $|z_1| \leq 1$ et $|z_2| \leq 1$, si et seulement si, les trois conditions suivantes sont satisfaites.*

$$P(z_1, 0) \neq 0 \text{ pour tout } z_1 \text{ tel que } |z_1| \leq 1,$$

$$P(1, z_2) \neq 0 \text{ pour tout } z_2 \text{ tel que } |z_2| \leq 1,$$

$$\mathbf{R}(z_1) \neq 0 \text{ pour tout } z_1 \text{ tel que } |z_1| = 1.$$

Remarque 1 Sur le cercle unité, $\overline{z}_1 = z_1^{-1}$, le résultant $\mathbf{R}(z_1)$, obtenu en éliminant z_2 entre les deux équations (2.5), peut alors être remplacé par celui obtenu en éliminant z_2 entre les deux équations :

$$P(z_1, z_2) = 0 \qquad \text{et} \qquad \tilde{P}(z_1, z_2) \stackrel{\Delta}{=} \sum_{h=0}^{n} \sum_{k=0}^{m} \overline{a}_{h,k} z_1^h z_2^k = 0.$$

Remarque 2 Le résultant $\mathbf{R}(z_1)$, obtenu en éliminant z_2 entre les deux équations (2.5), est égal, au signe près, au déterminant de la matrice de Schur-Cohn apparaissant dans les tests de Jury, Anderson, et Siljak. Quand le polynôme $P(z_1, z_2)$ est réel, pour $|z_1| = 1$, le résultant $\mathbf{R}(z_1)$ est un polynôme réel de la partie réelle de z_1. Le carré de ce polynôme divise le résultant apparaissant dans le premier test de Bose[7], et dans une version simplifiée[8], on retrouve ce polynôme au signe près. Dans le test de Maria et Fahmy, qui construit la table de Marden-Jury associée au polynôme (2.4) en z_2 dont les coefficients dépendent de z_1, l'unique élément de la dernière ligne de cette table est divisible par le résultant $\mathbf{R}(z_1)$.

Comme cela a déjà été remarqué par Fujiwara, le résultant $\mathbf{R}(z_1)$, obtenu en éliminant z_2 entre les deux équations (2.5), est égal, au signe près, au discriminant de la forme quadratique hermitienne de Schur-Cohn. Il est donc réel, quel que soit $z_1 \in \mathbb{C}$. Étant également une fonction polynôme complexe,

[7] "Implementation of a new stability test for two-dimensional filters", *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-25**, 117–120, 1977.

[8] "Simplification of a multidimensional digital filter stability test", *J. Franklin Institute*, **330**, *5*, 905–911, 1993.

F, des deux variables z_1 et $\overline{z}_1$, il existe un polynôme réel R à deux indéterminées, tel que pour tout nombre complexe $z_1 = x + iy$ décomposé suivant ses parties, réelle ou imaginaire,

$$\mathbf{R}(z_1) = F(z_1, \overline{z}_1) = R(x, y).$$

La division euclidienne du polynôme $R(x, y)$ par le polynôme $y^2 + x^2 - 1$ assure l'existence et l'unicité du polynôme réel S, aux deux indéterminées x, y et celles des deux polynômes réels Q et T en x, tels que :

$$\mathbf{R}(z_1) = R(x, y) = (y^2 + x^2 - 1)S(x, y) + yT(x) + Q(x). \tag{2.10}$$

Les polynômes R, S, Q et T dépendent évidemment des coefficients du polynôme P. Quand ce dernier est réel, on a

$$\overline{F(z_1, \overline{z}_1)} = F(\overline{z}_1, z_1),$$

donc $R(x, y) = R(x, -y)$ et T est identiquement nul.

Nous voyons sur l'équation (2.10), qu'il est plus simple de calculer le résultant $\mathbf{R}(z_1)$ quand z_1 est sur le cercle unité—ou sur une autre courbe algébrique— que quand z_1 est dans le disque unité fermé—ou dans un autre domaine de surface non nulle—. En effet, quand $|z_1| = 1$, il suffit de connaître les polynômes réels T et Q à une seule variable, pour évaluer $\mathbf{R}(z_1)$.

Comparons la complexité des différents critères de stabilité aux lumières de la proposition 7 et du point de vue exposé à sa suite, p. 20.

Tester si $P(z_1, z_2)$ s'annule sur le bi-disque $|z_1| \leq 1$ et $|z_2| \leq 1$, revient à examiner si $\mathbf{R}(z_1)$ s'annule sur le disque $|z_1| \leq 1$. Cela demande le calcul du polynôme réel $R(x, y)$ et le test de sa nullité sur le disque $x^2 + y^2 \leq 1$.

Tester si $P(z_1, z_2)$ s'annule sur le domaine $|z_1| = 1$ et $|z_2| \leq 1$, revient à vérifier si $\mathbf{R}(z_1)$ s'annule sur le cercle unité $|z_1| = 1$. Cela demande le calcul des polynômes réels Q et T qui n'ont qu'une indéterminée, et le test de nullité de $yT(x) + Q(x)$ sur le cercle $x^2 + y^2 = 1$, test que l'on sait faire en un nombre fini de pas. Pour cela, on élimine y entre les équations $yT(x) + Q(x) = 0$ et $x^2 + y^2 - 1 = 0$, pour obtenir l'équation algébrique en x :

$$(1 - x^2)\big(T(x)\big)^2 - \big(Q(x)\big)^2 = 0, \tag{2.11}$$

dont on sait calculer le nombre de racines réelles sur le segment $[-1; 1]$ grâce au théorème de Sturm.

Tester si $P(z_1, z_2)$ s'annule sur le bi-cercle $|z_1| = 1$ et $|z_2| = 1$, revient à examiner, si quand z_1 parcourt le cercle unité, le polynôme en z_2, $P(z_1, z_2)$, s'annule sur le cercle unité, c'est-à-dire s'il est sur la frontière de l'un des domaines $\mathcal{D}_p$ pour $0 \leq p \leq n$ ou autrement dit dans $\mathcal{B}_1$, d'après la proposition 2. Pour vérifier cette condition, il faut tester si $\mathbf{R}(z_1)$ s'annule sur le cercle unité, puis dans le cas où $\mathbf{R}(z_1)$ s'y annulerait, il faut s'assurer que ce n'est pas en des points de $\mathcal{B}_2 \setminus \mathcal{B}_1$. Vu sous cet angle, il apparaît que le critère de Strintzis n'est pas plus facile à tester que celui de Huang-Goodman.

Nous venons de voir que le test de nullité du résultant $\mathbf{R}(z_1) = R(x, y)$ sur le disque unité $x^2 + y^2 \leq 1$, peut se simplifier en se ramenant au test de nullité de R modulo $y^2 + x^2 - 1$:

$$R(x, y) \equiv yT(x) + Q(x) \qquad (y^2 + x^2 - 1)$$

sur le cercle unité. Nous nous sommes alors posé la question de savoir si l'on peut encore réduire le domaine du plan sur lequel on teste la nullité du résultant $R(x, y)$ et nous avons démontré que cela est impossible (voir le chapitre II du rapport de thèse).

Nous avons vu au paragraphe précédent que tout algorithme, qui décide en un nombre fini de pas si le polynôme $P(z_1, z_2)$ s'annule sur le bi-disque unité fermé, revient, à un simple test de stabilité 1-D près, à décider si le résultant $\mathbf{R}(z_1)$ s'annule sur un domaine du plan. Tous les critères de stabilité que nous venons d'examiner reviennent à tester si le résultant $\mathbf{R}(z_1) = R(x, y)$, où $z_1 = x + iy$, s'annule sur un sous-ensemble Δ du plan, qui vaut soit le disque unité $y^2 + x^2 \leq 1$, soit le cercle unité.

Nous avons défini la *complexité* de ces critères comme étant égale au nombre de coefficients du polynôme R qu'il faut connaître pour calculer $R(x, y)$ sur Δ. Avec cette définition, nous n'avons pas pris en compte la complexité du test de nullité de R, car elle dépend de l'algorithme utilisé et rien ne dit qu'il n'en existe pas de bien meilleurs que ceux connus aujourd'hui. Toutefois, nous avons admis que pour tester en un nombre fini de pas la nullité de $R(x, y)$ sur Δ, il faut connaître ses coefficients.

Si P est de degré n en z_1 et m en z_2, alors R est de degré nm en x et en y. Tester si R s'annule sur le disque $y^2 + x^2 \leq 1$ a une complexité de $(nm + 1)^2$. Tester si R s'annule sur le cercle unité a une complexité de $2(nm + 1)$ dans le cas complexe (coefficients de T et Q), et de $nm + 1$ dans le cas réel. Cette complexité est minimale, aucun critère de stabilité vérifiable en un nombre fini de pas ne peut avoir une complexité moindre. Notre critère, comme celui de Huang-Goodman, est de complexité minimale.

Intéressons-nous maintenant au cas où le polynôme P, défini en (2.1), est réel. Remarquons d'abord, qu'en vertu de la proposition 5, le résultant $\mathbf{R}(z_1)$ associé ne se factorise pas en général, car tout polynôme réel P, défini en (2.1) et considéré comme un polynôme en z_2, a des coefficients, qui peuvent *a priori* prendre n'importe quelle valeur complexe, quand z_1 est sur le cercle unité. Quand P est réel, de degré n en z_1 et m en z_2, nous avons vu que le reste de la division du résultant $R(x, y)$ par $y^2 + x^2 - 1$ se réduit à un polynôme réel Q en x, de degré nm.

Remarque 3 Les résultats de cette section, relatifs à la stabilité de filtres bi-dimensionnels quart-plans causals, peuvent s'étendre sans aucune difficulté aux filtres demi-plans semi-causals.

2.3 Deux algorithmes rapides pour tester la stabilité de filtres numériques récursifs 2-D

Un algorithme rapide pour tester la stabilité de filtres numériques récursifs bi-dimensionnels quart-plans causaux et dépourvus de toute singularité non essentielle de la deuxième espèce est décrit dans cette section, c'est un résultat de ma thèse. Pour traiter des filtres, dont le dénominateur de la fonction de transfert est un polynôme réel

$$P(z_1, z_2) = \sum_{h=0}^{n} \sum_{k=0}^{m} a_{h,k} z_1^{n-h} z_2^{m-k} = \sum_{k=0}^{m} a_k(z_1) z_2^{m-k} \quad (m \leq n) \tag{2.12}$$

de degré n en z_1 et m en z_2, il demande $(21/2)m^2n^2 + 36m^3n + O(mn^2)$ opérations arithmétiques. Parmi les algorithmes équivalents connus en 1993, les plus rapides requéraient, quand $m = n$, $O(n^5)$ opérations soit un ordre de grandeur de plus que celui présenté ici.

Ce dernier utilise le résultat suivant (apparaissant dans la thèse et publié avec l'algorithme complet dans [11]). La matrice de Shur-Cohn[9] construite à partir du polynôme complexe de degré n

$$P = a_0 x^n + \cdots + a_n \tag{2.13}$$

[9]définie par

$$\mathbf{T} = \left(\mathbf{A}^{\bullet}\right)^{H} \mathbf{A}^{\bullet} - \mathbf{A}^{H} \mathbf{A}$$

a un rang de déplacement[10] égal à deux en général. Il en résulte que l'algorithme de Levinson-Szegö généralisé[11] peut être employé pour calculer, en $O(n^2)$ opérations, la signature de la forme quadratique hermitienne associée à cette matrice et, en particulier, le résultant de P et de P^*, défini à la relation (2.2). Calculer le résultant de deux polynômes de degré n en $O(n^2)$ opérations n'est pas nouveau. La table de Routh (par exemple) permet de le faire. Utiliser à cet effet le programme de Levinson-Szegö généralisé—qu'il est très facile d'étendre à des matrices hermitiennes complexes—n'avait toutefois encore jamais été fait.

Il a été indiqué à la section précédente que tout algorithme décidant si un polynôme (2.12) admet des zéros dans le bi-disque $|z_1| \leq 1$ et $|z_2| \leq 1$, doit nécessairement tester si le résultant $\mathbf{R}(z_1)$, obtenu en éliminant z_2 entre les deux équations (2.5), s'annule sur le cercle unité $|z_1| = 1$. Effectivement, les tests classiques de Maria et Fahmy, Anderson Jury et Siljak, Kanellakis Tzafestas et Theodorou (voir la bibliographie de [48] pour trouver les articles associés), dérivés de la table de Schur-Cohn-Jury, calculent explicitement ce résultant (ou un polynôme divisible par ce résultant) sous la forme d'une fonction polynôme en z_1 et $\overline{z}_1$; et le programme le plus rapide, selon O'Connor et Huang, nécessite $O[mn(nm^2 + m^3)]$ opérations arithmétiques.

Plus récemment, toujours basé sur la table de Shur-Cohn-Jury, la méthode de Hu[12] évalue, en $O(4^m mn^2)$ opérations, le résultant $\mathbf{R}(z_1)$ ci-dessus, uniquement sur le cercle unité ($\overline{z}_1 = z_1^{-1}$).

L'algorithme de Gu et Lee[13] calcule $\mathbf{R}(z_1)$ sur le cercle $|z_1| = 1$ par interpolation, en donnant à z_1, $nm + 1$ valeurs particulières ζ_k ($0 \leq k \leq mn$) sur ce cercle. Il évalue, pour chacun de ces points, $\mathbf{R}(\zeta_k)$ avec l'algorithme de factorisation de Cholesky en $O(m^3)$ opérations. Cette étape est la plus complexe de leur programme et requiert $O(m^4 n)$ opérations.

La méthode que nous avons proposée calcule, pour un polynôme (2.12) réel, le reste $Q(x)$—défini à la relation (2.10)—de la division de $R(x, y) = \mathbf{R}(z_1)$ (où $z_1 = x+iy$ est décomposé suivant ses parties, réelle ou imaginaire) par $y^2 + x^2 - 1$. Les coefficients de Q, dont le degré vaut nm, sont obtenus par interpolation—comme dans le test de G. Gu et E. B. Lee—après avoir évalué $\mathbf{R}(z_1)$ en $nm+1$ points du cercle unité, au moyen de l'algorithme de Levinson-Szegö généralisé et en $O(m^2)$ opérations. Cette étape du programme a une complexité en $O(nm^3)$ et le calcul de Q demande $O(n^2 m^2)$ opérations, tout comme le test complet. Elle est décrite dans le rapport de thèse ou le livre [1] pages 226–231.

Après la thèse, j'ai développé un autre algorithme, il est décrit dans [1] aux pages 231–235 (ou dans [3], pp. 372–376). Il suppose que le polynôme (2.12) est réel et que le plus petit des degrés n ou m est inférieur ou égal à 4, le plus grand étant quelconque. Ce dernier algorithme se révèle être très robuste face aux erreurs d'arrondi comme nous allons le voir à la section suivante.

---

où pour une matrice $\mathbf{X}$, sa matrice transposée conjuguée est notée $\mathbf{X}^H$, et

$$\mathbf{A} = \begin{bmatrix} a_n & a_{n-1} & \cdots & a_1 \\ 0 & a_n & \cdots & a_2 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_n \end{bmatrix} \qquad \text{et} \qquad \mathbf{A}^\bullet = \begin{bmatrix} \overline{a}_0 & \overline{a}_1 & \cdots & \overline{a}_{n-1} \\ 0 & \overline{a}_0 & \cdots & \overline{a}_{n-2} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \overline{a}_0 \end{bmatrix}.$$

[10] Voir l'article de Kailath, Kung et Morf, "Displacement ranks of a matrix", *Bull. of the Amer. Math. Soc.*, **1**, *5*, 769–773, 1979.

[11] Voir l'article de Friedlander, Kailath, Morf et Ljung, "Extended Levinson and Chandrasekhar equations for general discrete-time linear estimation problems", *IEEE Trans. on Aut. and Cont.*, **AC-23**, *4*, 653–659, 1978.

[12] "2-D Filter Stability Tests Using Polynomial Array for $F(z_1, z_2)$ on $|z_1| = 1$", *IEEE Tran. on Circuits and Systems*, **38**, *9*, 1092–1095, 1991.

[13] "A Numerical Algorithm for Stability Testing of 2-D Recursive Digital Filters", *IEEE Trans. on Circ. and Syst.*, **37**, *1*, 135–138, 1990.

2.4 Comparaison des algorithmes face aux erreurs d'arrondi

Dans cette section nous présentons des résultats, publiés dans [9], relatifs au comportement de quatre algorithmes testant la stabilité de filtres récursifs 2D face aux erreurs d'arrondi quand ils sont implantés avec une arithmétique à virgule flottante. Parmi ces quatre algorithmes figurent les deux de la section précédente (que nous noterons (BB1) pour le premier et (BB2) pour le second auxquels ont été ajoutés un de Hu (celui mentionné ci-dessus) (que nous noterons (H)) et une version améliorée que Hu a réalisée avec la collaboration de Jury[14] (notée (HJ)). Considérons un filtre numérique récursif 2D causal, dépourvu de toute singularité non essentielle de la deuxième espèce sur le bi-cercle unité et dont le dénominateur de la fonction de transfert est le polynôme **réel**

$$P(z_1, z_2) = \sum_{h=0}^{n} \sum_{k=0}^{m} a_{h,k} z_1^{n-h} z_2^{m-k} = \sum_{k=0}^{m} a_k(z_1) z_2^{m-k} \quad (m \leq n) \tag{2.14}$$

et notons $\mathbf{R}(z_1)$ le résultant obtenu en éliminant z_2 entre les équations

$$P(z_1, z_2) = 0 \quad \text{et} \quad P^\star(z_1, z_2) = \sum_{k=0}^{m} \overline{a_k(z_1)} z_2^k. \tag{2.15}$$

Le résultant $\mathbf{R}(z_1)$ est un polynôme réel des variables z_1 et $\overline{z}_1$ dont les coefficients s'expriment au moyen de polynômes à coefficients entiers dont les variables sont les coefficients $a_{h,k}$ du polynôme (2.14). Les algorithmes (H) et (HJ) étant basés sur la table de Jury 1D, commençons par rappeler la procédure de sa construction. Considérons un polynôme complexe à une variable $P(z) = a_m z^m + a_{m-1} z^{m-1} + \ldots + a_0$, on construit, ligne par ligne, une table à $m+1$ lignes et colonnes, numérotées de 0 à m. L'élément situé à l'intersection de la ligne i et de la colonne j ($0 \leq i \leq m$ et $0 \leq j \leq m-i$), noté $a_{i,j}$, se calcule avec les équations $a_{0,j} = a_j$ pour $0 \leq j \leq m$ et

$$a_{i,j} = \begin{vmatrix} a_{i-1,0} & a_{i-1,m-i-j+1} \\ \overline{a}_{i-1,m-i+1} & \overline{a}_{i-1,j} \end{vmatrix} \tag{2.16}$$

pour $i = 1, 2, \ldots, m$ et $j = 0, 1, \ldots, m-i$. L'algorithme de Hu étend la construction de la table de Jury au cas où les éléments sont des polynômes en z_1 et $\overline{z}_1$, il calcule ainsi, pour $|z_1| = 1$, un polynôme réel $S(z_1, \overline{z}_1)$ symétrique[15] dont le degré par rapport à chacune des variables vaut $n2^{m-1}$. Ce polynôme correspond au dernier élément construit avec la table. L'algorithme (H) n'est pas optimal car le polynôme $S(z_1, \overline{z}_1)$ est strictement divisible par le résultant $\mathbf{R}(z_1)$. Pour corriger ce défaut, Hu et Jury ont proposé l'algorithme (HJ) qui est basé sur une table de Jury 1D construite comme ci-dessus, l'équation (2.16) étant remplacée par

$$a_{i,j} = \frac{1}{c_i} \begin{vmatrix} a_{i-1,0} & a_{i-1,m-i-j+1} \\ \overline{a}_{i-1,m-i+1} & \overline{a}_{i-1,j} \end{vmatrix} \tag{2.17}$$

pour $i = 1, 2, \ldots, m$ et $j = 0, 1, \ldots, m-i$, où

$$c_i = \begin{cases} 1 & \text{si } i \leq 2 \\ a_{i-2,0} & \text{si } i > 2. \end{cases} \tag{2.18}$$

Le dernier élément de la table ainsi calculée, quand la première ligne est constituée des polynômes en z_1 correspondant aux coefficients de $P(z_1, z_2)$ considéré comme un polynôme en z_2, est le résultant $\mathbf{R}(z_1)$.

Les quatre algorithmes passent tous par les trois étapes suivantes :

[14] "On two-dimensional filter stability test", *IEEE Trans. Circuits and Systems*, **41**, 457–462, 1994.

[15] C'est-à-dire satisfaisant à l'identité $S(z_1, \overline{z}_1) = S(\overline{z}_1, z_1)$.

1. Pour $|z_1| = 1$, calculer le résultant $\mathbf{R}(z_1)$ ou un autre polynôme réel en z_1 et $\overline{z}_1$: $S(z_1, \overline{z}_1)$.
2. Transformer le résultant $\mathbf{R}(z_1)$ ou le polynôme $S(z_1, \overline{z}_1)$ en un polynôme réel à une seule variable $Q(x)$, où $x/2$ est la partie réelle de z_1.
3. À l'aide du théorème de Sturm, tester si le polynôme $Q(x)$ s'annule sur le segment réel $[-2\,,\,2]$.

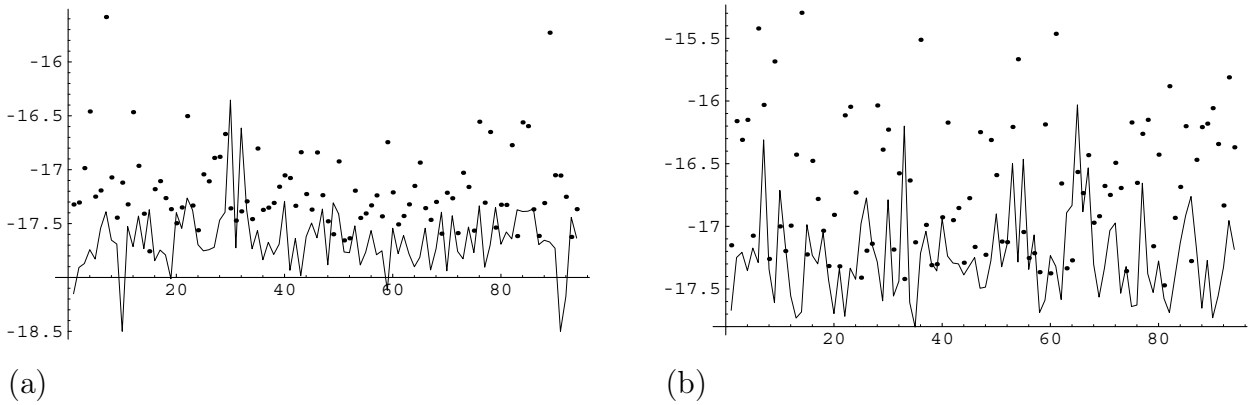
Nous avons implanté les quatre algorithmes sur un ordinateur PC avec le processeur 80 386 et le sytème d'exploitation MSDOS, utilisant le compilateur Turbo C 2.0 et des données flottantes codées par 10 mots de 8 bits chacune. Ce qui correspond à la triple précision : 64 bits sont réservés à la mantisse et 15 à l'exposant. Pour étudier les erreurs d'arrondi des algorithmes, nous les avons également traduits en des procédures du logiciel de calcul formel `Mathematica`, avec lequel on peut mener des opérations arithmétiques avec une précision infinie quand les données sont des nombres rationnels. Nous avons ainsi calculé exactement les coefficients du polynôme $S(z_1, \overline{z}_1)$, ceux du résultant $\mathbf{R}(z_1)$ et ceux du polynôme $Q(x)$ ci-dessus. Nous avons généré un grand nombre de polynômes (2.14) à coefficients entiers choisis au hasard avec une densité de probabilité uniforme sur le segment $[-100\,,\,100]$. Pour chaque polynôme nous avons comparé les résultats donnés par l'algorithme implanté sur le processeur à virgule flottante à ceux obtenus avec `Mathematica`. Pour des degrés n et m du polynôme (2.14) fixés, nous avons testé 94 polynômes différents, numérotés de 1 à 94. Cet indice sera désigné par *le numéro de tirage du polynôme* dans les figures ci-dessous. Puis nous avons fait varier les degrés. Voici les résultats que nous avons obtenus.

2.4.1 Robustesse des quatre algorithmes testés

Pour l'algorithme (H) nous avons estimé l'erreur d'arrondi relative faite sur les coefficients du polynôme $S(z_1, \overline{z}_1)$. De façon précise, notons, pour $0 \leq i \leq n2^{m-1}$, α_i, la valeur exacte du coefficient de degré i en z_1 (ou $\overline{z}_1$) du polynôme $S(z_1, \overline{z}_1)$ et $\alpha_{F,i}$ celle calculée au moyen de l'algorithme (H) avec une arithmétique à virgule flottante. La quantité

$$\varepsilon(S) = \max_{0 \leq i \leq n2^{m-1}} \left| \frac{\alpha_{F,i} - \alpha_i}{\alpha_i} \right| \tag{2.19}$$

correspond à l'erreur d'arrondi relative faite par l'algorithme (H) pour calculer les coefficients de $S(z_1, \overline{z}_1)$. La figure 2.1 donne les résultats obtenus pour différentes valeurs de n et m. La quantité



(a) (b)

FIG. 2.1 – $\log_{10} \varepsilon(S)$ pour (a) $n = m = 4$ (trait fin), $n = 15$, $m = 4$ (pointillés) et (b) $n = m = 5$ (trait fin), $n = 15$, $m = 5$ (pointillés).

$\log_{10}\varepsilon(S)$ et le numéro de tirage du polynôme sont représentés avec une échelle linéaire sur les axes respectifs Y et X.

Pour les algorithmes (HJ) et (BB2), nous avons estimé l'erreur d'arrondi relative faite sur les $nm+1$ coefficients du résultant $\mathbf{R}(z_1)$, en définissant de façon similaire la quantité

$$\varepsilon(\mathbf{R}) = \max_{0\leq i\leq nm} \left| \frac{\beta_{F,i} - \beta_i}{\beta_i} \right|,$$

où la grandeur $\beta_{F,i}$ (resp. β_i) désigne le coefficient de degré i du résultant calculé avec une arithmétique à virgule flottante (resp. une arithmétique rationnelle exacte). La figure 2.2 montre les résultats que

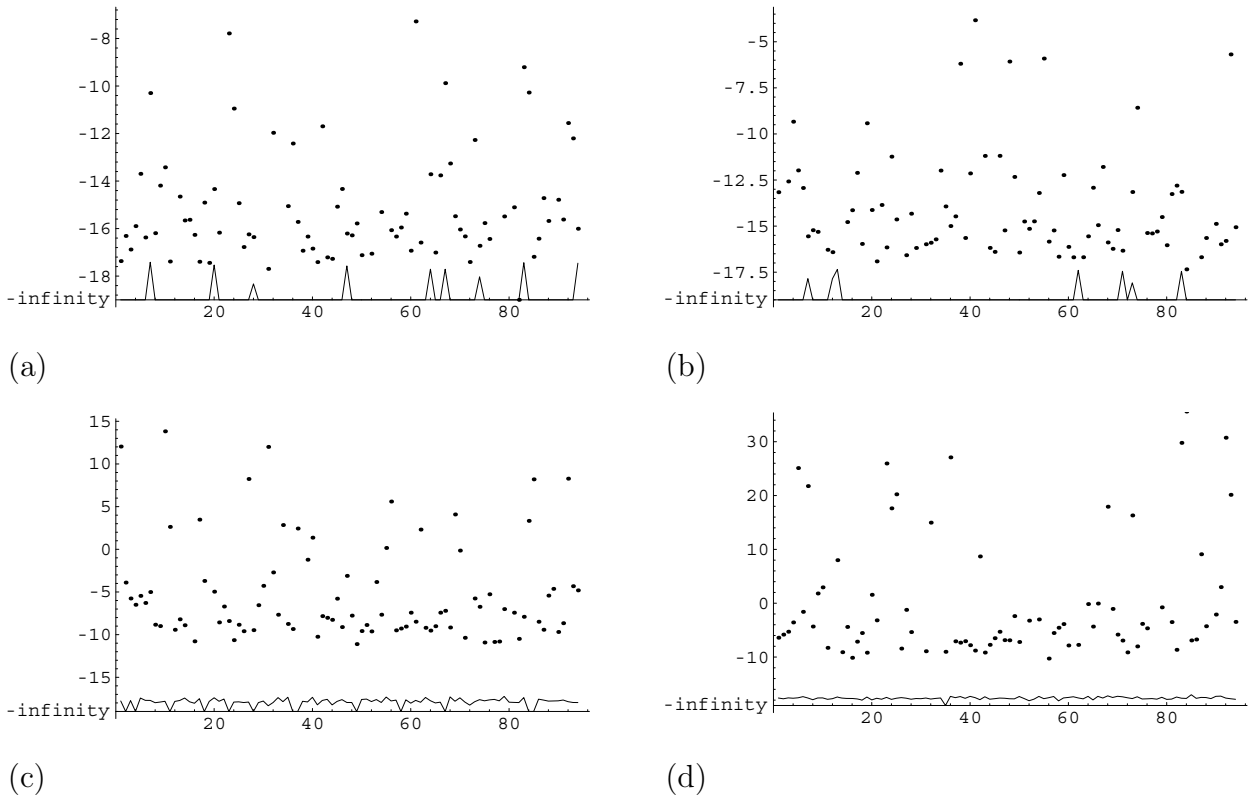


FIG. 2.2 – $\log_{10}\varepsilon(\mathbf{R})$ pour (a) $n=10$, $m=3$, $N_d=8$, (b) $n=15$, $m=3$, $N_d=9$, (c) $n=m=4$, $N_d=3$ et (d) $n=10$, $m=4$, $N_d=16$ avec (BB2) (trait fin) et (HJ).

nous avons obtenus pour différentes valeurs de n et m. Pour garder une lecture agréable des courbes, quelques points, associés à des valeurs trop grandes pour l'échelle, n'ont pas été tracés. Pour chacune des courbes, nous indiquons le nombre N_d de ces points quand il est non nul. La quantité $\log_{10}\varepsilon(\mathbf{R})$ et le numéro de tirage du polynôme apparaissent en échelle linéaire sur les axes respectifs Y et X.

De plus, pour les algorithmes (HJ), (BB1) et (BB2), nous avons estimé l'erreur d'arrondi relative faite sur les $mn+1$ coefficients du polynôme $Q(x)$ introduit ci-dessus. Nous avons défini la quantité $\varepsilon(Q)$ associée au polynôme Q de la même façon que la quantité (2.19) est associée au polynôme $S(z_1,\overline{z}_1)$. Les figures 2.3 et 2.4 montrent les résultats obtenus pour différentes valeurs de n et m. La quantité $\log_{10}\varepsilon(Q)$ et le numéro de tirage du polynôme apparaissent en échelle linéaire sur les axes respectifs Y et X.

Pour tester si le polynôme $Q(x)$ s'annule sur le segment réel $[-2\,,\,2]$ à l'aide du théorème de Sturm, on procède de la manière suivante. Partant de $f_0(x)=Q(x)$ et $f_1(x)=Q'(x)$, le polynôme

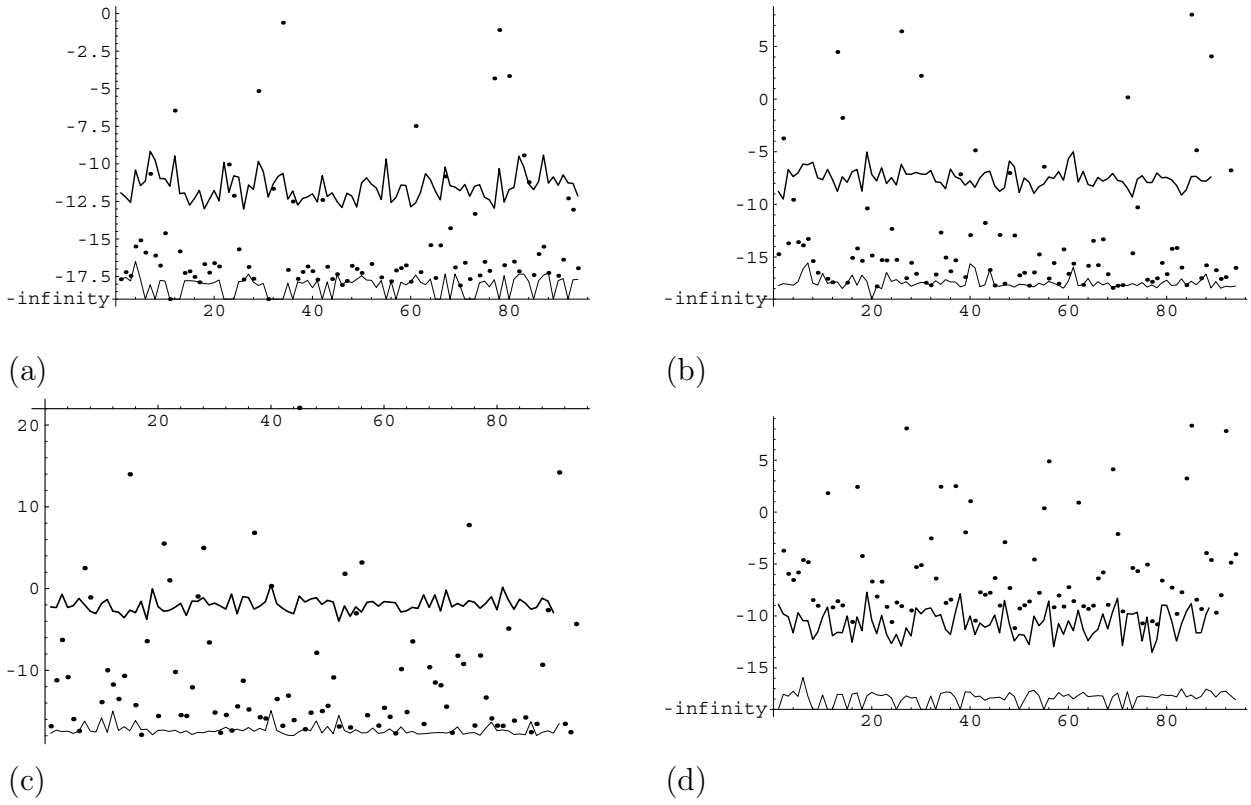


FIG. 2.3 – $\log_{10}\varepsilon(Q)$ pour (a) $n = 10$, $m = 3$, $N_d = 1$, (b) $n = 15$, $m = 3$, (c) $n = 20$, $m = 3$, $N_d = 4$ et (d) $n = m = 4$, $N_d = 5$ avec (BB2) (trait fin), (BB1) (trait épais) et (HJ) (pointillés).

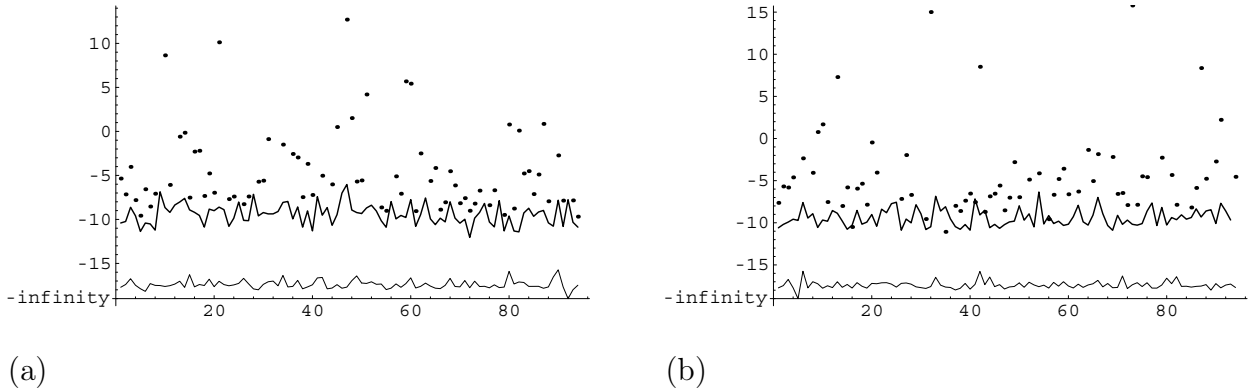


FIG. 2.4 – $\log_{10}\varepsilon(Q)$ pour (a) $n = 8$, $m = 4$, $N_d = 19$ et (b) $n = 10$, $m = 4$, $N_d = 26$ avec (BB2) (trait fin), (BB1) (trait épais) et (HJ) (pointillés).

dérivé de Q, on calcule la famille des polynômes

$$f_0(x), f_1(x), \ldots, f_p(x), \ldots, f_r(x) \tag{2.20}$$

par divisions euclidiennes successives et changement de signe du reste : $-f_{p+1}(x)$ est le reste de la division euclidienne de $f_{p-1}(x)$ par $f_p(x)$, jusqu'à ce que $f_{r+1}(x)$ soit identiquement nul. En notant $V(x)$ la famille (2.20), le nombre de racines réelles distinctes de $Q(x)$ dans le segment réel $[-2\,,\,2]$ s'obtient aisément à partir des changements de signes — les termes nuls n'étant pas pris en compte — des suites $V(-2)$ et $V(2)$. Pour les algorithmes (HJ), (BB1) et (BB2) nous avons estimé l'erreur d'arrondi relative faite sur le calcul des termes de $V(-2)$ et $V(2)$ par une quantité similaire à celle donnée par la relation (2.19) et notée $\varepsilon(V)$. Pour éviter une trop grande croissance des coefficients des

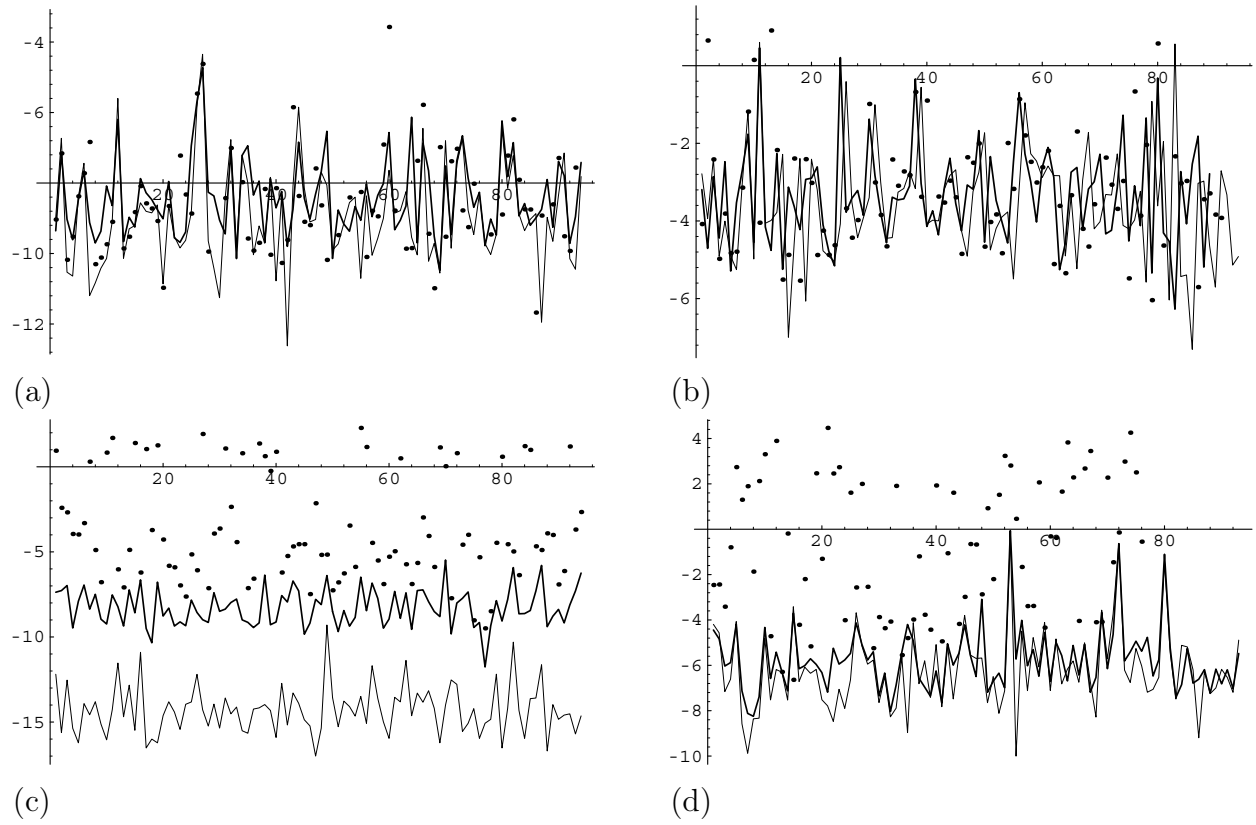


FIG. 2.5 – $\log_{10}\varepsilon(V)$ pour (a) $n = 10$, $m = 3$, $N_d = 11$, (b) $n = 15$, $m = 3$, $N_d = 7$, (c) $n = m = 4$ et (d) $n = 10$, $m = 4$ avec (BB2) (trait fin), (BB1) (trait épais) et (HJ) (pointillés).

polynômes de la famille (2.20), nous les avons "normalisés" de la façon suivante. Avant chaque division euclidienne, le diviseur est normalisé, c'est-à-dire que, premièrement, son plus grand coefficient (en valeur absolue) est recherché puis chacun de ses coefficients est divisé par cette plus grande valeur absolue. Si le résultat de la division est inférieur, en valeur absolue, à 10^{-32}, alors le coefficient est mis à zéro. Après cette opération de normalisation, le nouveau polynôme obtenu a tous ses coefficients inférieurs, en valeur absolue, à 1 et certains d'entre eux, petits, ont pu être forcés à zéro[16]. Nous avons constaté que cette "normalisation" du diviseur améliorait sensiblement le comportement face

[16]Lors de l'implantation des algorithmes avec `Mathematica`, nous avons également normalisé les polynômes de telle sorte, qu'en valeur absolue, tous leurs coefficients soient inférieurs à 1, mais aucun coefficient n'est forcé à zéro (même si sa valeur absolue est inférieure à 10^{-32}.)

aux erreurs d'arrondi des algorithmes. Dans certains cas, les erreurs d'arrondi sur les coefficients du polynôme Q sont si importantes que la famille des polynômes (2.20) n'a pas le même nombre de termes quand elle est calculée en flottant ou avec une précision infinie. Dans ce cas, l'erreur relative $\varepsilon(V)$ ne peut pas être mesurée, c'est pourquoi quelques unes des courbes suivantes ne sont pas complètes. La figure 2.5 montre les résultats obtenus pour différentes valeurs de n et m. La quantité $\log_{10}\varepsilon(V)$ et le numéro de tirage du polynôme apparaissent en échelle linéaire sur les axes respectifs Y et X.

Enfin, pour les algorithmes (HJ), (BB1) et (BB2) nous avons compté le nombre d'erreurs de signe dans les familles $V(-2)$ et $V(2)$, c'est la quantité $ES(V)$. La figure 2.6 montre les résultats que nous avons obtenus pour différentes valeurs de n et m. La quantité $SE(V)$ et le numéro de tirage du polynôme apparaissent en échelle linéaire sur les axes respectifs Y et X.

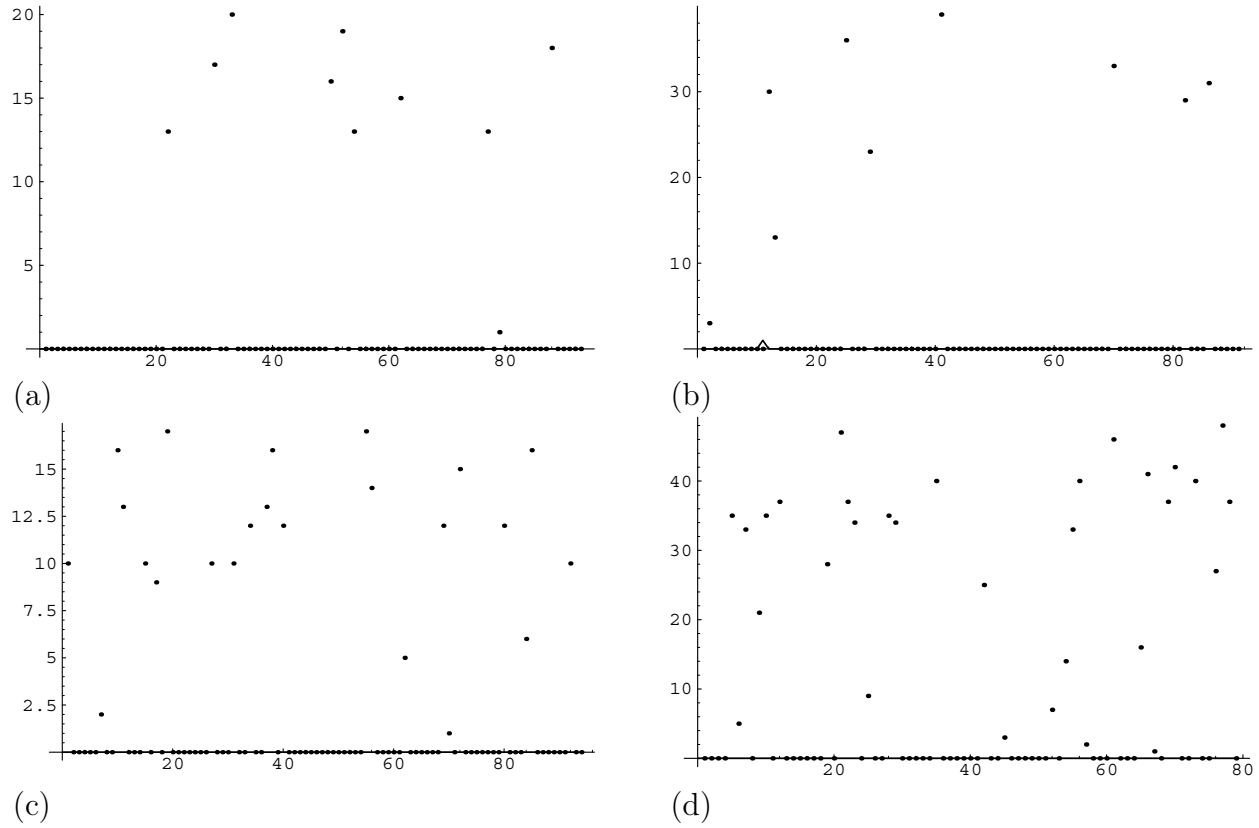


FIG. 2.6 – $SE(V)$ pour (a) $n = 10$, $m = 3$, (b) $n = 15$, $m = 3$, (c) $n = m = 4$ et (d) $n = 10$, $m = 4$, avec (BB2) (trait fin), (BB1) (trait épais) et (HJ) (pointillés).

2.4.2 Analyse des différences de comportement

Nous pouvons constater une différence notable de comportement face aux erreurs d'arrondi des algorithmes (HJ), (BB1) et (BB2). L'algorithme (BB2) est sensiblement meilleur que les autres. En revanche, il est plus difficile à implanter sur un ordinateur. Le temps de calcul requis par chacun des quatre algorithmes est à peu près le même pour de faibles valeurs du produit nm ($nm \leq 40$). On constate également qu'avec l'algorithme (HJ), pour des degrés n et m fixés, le crédit que l'on peut accorder à la réponse et aux résultats intermédiaires varie considérablement d'un polynôme (2.14) à l'autre. Même pour de faibles valeurs du plus petit des degrés m, la réponse de l'algorithme (HJ)

peut être fausse à cause des erreurs d'arrondi. Ceci est dû à l'introduction, lors de la construction de la table de Jury-Hu, (cf les équations (2.17) et (2.18)) de la division euclidienne par les quantités c_i $(0 \leq i \leq m)$ (qui sont des polynômes symétriques en z_1 et $\overline{z}_1$), afin de réduire les degrés des polynômes $a_{i,j}$ apparaissant dans l'algorithme (H), et rendre ainsi son degré de complexité minimal. Ce mauvais comportement de l'algorithme (HJ) n'apparaît pas avec l'algorithme (H), où aucune division de polynômes n'est effectuée.

L'algorithme de division euclidienne de polynômes est très peu robuste aux erreurs d'arrondi, car il arrive souvent que, lors des calculs intermédiaires, deux grands flottants soient retranchés l'un à l'autre, donnant un résultat encore grand (pour l'exposant) mais avec une faible mantisse, augmentant ainsi, en une seule instruction et de façon très importante, l'erreur d'arrondi relative. Ce défaut apparaît très souvent quand les divisions euclidiennes s'enchaînent et quand les coefficients du numérateur et du dénominateur ne sont pas normalisés (c'est-à-dire quand ils peuvent prendre des valeurs de plus en plus grandes). Lors de la description de l'implantation effective de l'algorithme de Sturm nous avons présenté une méthode de normalisation des polynômes (2.20) qui ne s'applique pas aux divisions euclidiennes de l'algorithme (HJ) : il est possible de normaliser, avec notre méthode, les dénominateurs successifs c_i $(0 \leq i \leq m)$, mais il est impossible de normaliser tous les numérateurs car le même c_i divise plusieurs numérateurs différents. Nous avons également constaté, pour tous les algorithmes, que tant que l'erreur d'arrondi relative (au fur et à mesure que les instructions se déroulent) est suffisamment petite, elle reste négligeable jusqu'à la fin de la procédure, nous disons dans ces conditions que l'algorithme est *robuste*. En revanche, dès qu'elle dépasse un seuil, elle devient tout de suite très grande, en très peu d'instructions, nous disons alors que l'algorithme *explose*.

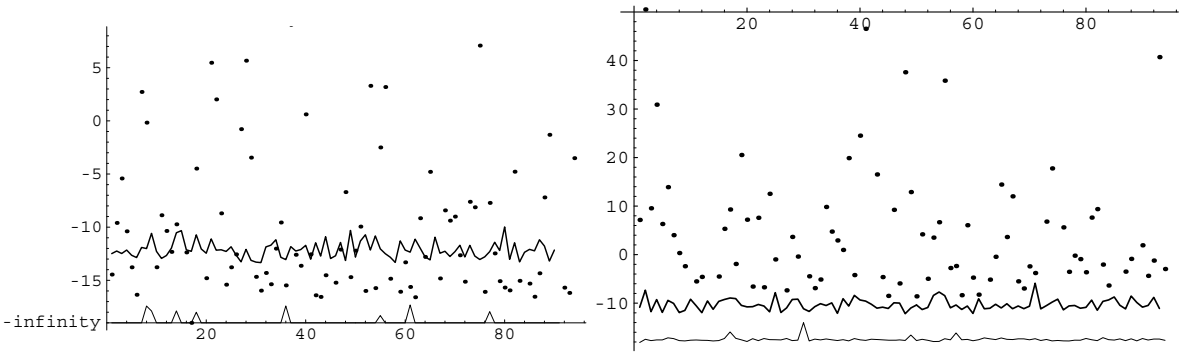
Le lecteur peut se demander pourquoi l'algorithme (BB2) semble toujours donner une bonne réponse avec des erreurs d'arrondi relatives quasi nulles sur les calculs intermédiaires, pour les degrés satisfaisant aux inégalités $m \leq 4$ et $mn \leq 40$. Indiquons que, sans aucun doute, il existe des polynômes (2.14) satisfaisants aux conditions précédentes sur les degrés pour lesquels la réponse de l'algorithme (BB2) est fausse. Ce n'est pas parce que nous n'en avons pas rencontrés en les tirant au hasard qu'ils n'existent pas. Pour la même raison, il en existe peut-être dont les coefficients sont dans l'intervalle $[-100\,,\,+100]$. Nous pouvons simplement affirmer que nous n'en avons jamais rencontrés malgré le grand nombre de tests réalisés. Pour calculer le résultant $\mathbf{R}_2(z_1)$ obtenu en éliminant z_2 entre les équations

$$P(z_1, z_2) = 0 \quad \text{et} \quad \tilde{P}(z_1, z_2) = \sum_{h=0}^{n} \sum_{k=0}^{m} a_{h,k} z_1^h z_2^k = 0,$$

l'algorithme (BB2) développe le déterminant de la matrice de Bézout $\mathbf{M}(z_1)$ dont les éléments sont des polynômes en la seule variable z_1. Lors du traitement de ce développement, l'algorithme utilise le plus grand nombre possible des symétries entre les termes des calculs intermédiaires. Pour éviter de calculer plusieurs fois les mêmes coefficients, il force les symétries des résultats intermédiaires. Cette stratégie permet de calculer les coefficients du résultant $\mathbf{R}_2(z_1)$ avec une erreur d'arrondi relative suffisamment faible pour que, après la transformation linéaire (très mal conditionnée) qui associe le polynôme réel $Q(x)$ au polynôme $\mathbf{R}_2(z_1)$, l'erreur d'arrondi relative sur les coefficients de Q reste assez petite et qu'en fin de compte, après la procédure basée sur le théorème de Sturm et la normalisation des polynômes intermédiaires, le résultat final soit juste. Pour implanter l'algorithme (HJ), nous avons également adopté la stratégie de forcer au maximum les symétries des résultats intermédiaires. Cependant, elle n'est pas très efficace ici : elle retarde de quelques instructions le moment à partir duquel l'algorithme explose, mais ne permet pas de maintenir l'erreur d'arrondi relative en dessous du seuil de robustesse. Les symétries qui apparaissent naturellement dans les éléments de la table de Jury-Hu (2.17) et (2.18) sont beaucoup moins nombreuses que celles exploitées par l'algorithme

(BB2).

Dans l'algorithme (BB1), une première étape calcule la transformée de Fourier discrète (TFD) sur $2(nm+1)$ points des coefficients du résultant $\mathbf{R}(z_1)$ au moyen de l'algorithme de Levinson-Szegö généralisé. Par TFD inverse, il est facile de calculer les coefficients de $\mathbf{R}(z_1)$. Nous avons alors estimé pour l'algorithme (BB1) l'erreur d'arrondi relative sur les coefficients de $\mathbf{R}(z_1)$. La figure 2.7 montre les résultats obtenus. La quantité $\log_{10}\varepsilon(\mathbf{R})$ et le numéro de tirage du polynôme apparaissent en échelle linéaire sur les axes respectifs Y et X.



(a) (b)

FIG. 2.7 – $\varepsilon(\mathbf{R})$ pour (a) $n=20$, $m=3$, $N_d=8$ et (b) $n=15$, $m=4$, $N_d=5$ avec (BB2) (trait fin), (BB1) (trait épais) et (HJ) (pointillés).

Nous remarquons que l'algorithme (BB1) obtient une relativement faible erreur d'arrondi relative sur les coefficients de $\mathbf{R}(z_1)$, de l'ordre de 10^{-12} à 10^{-10}. Cependant, cette erreur est déjà trop grande pour assurer la robustesse de la transformation qui à $\mathbf{R}(z_1)$ associe $Q(x)$ (quand $nm=60$), car cette transformation est trop mal conditionnée.

D'autre part, le calcul du dernier élément $S(z_1,\overline{z}_1)$ de la table de Jury par l'algorithme (H) résiste bien aux erreurs d'arrondi. Mais les deux dernières étapes (calcul du polynôme $Q(x)$ associé à $S(z_1,\overline{z}_1)$ et celui du nombre de zéros de $Q(x)$ dans l'intervalle $[-2\,,\,2]$ avec la méthode de Sturm) ne sont pas aussi bonnes. Voici trois raisons à cela. La première vient d'être mentionnée à propos du mauvais conditionnement de la transformation qui associe $Q(x)$ à $S(z_1,\overline{z}_1)$, le conditionnement se dégradant quand les degrés de $S(z_1,\overline{z}_1)$ augmentent. La deuxième vient du fait que les perturbations sur la localisation des zéros d'un polynôme, quand les coefficients sont soumis à de faibles perturbations, augmentent en général avec le degré du polynôme[17]. Enfin, la troisième est que le nombre de termes dans la suite (2.20) est nettement plus important dans l'algorithme (H) en comparaison avec les trois autres.

2.4.3 Conclusion

Nous recommandons l'emploi du deuxième algorithme quand cela est possible, c'est-à-dire quand le plus petit des degré du polynôme apparaissant au dénominateur du filtre récursif 2D satisfait à la condition $m \leq 4$. En général dans les applications en traitement d'images cette contrainte est satisfaite. Dans le cas contraire, lors d'une implantation avec une arithmétique à virgule flottante, quel que soit l'algorithme utilisé, il ne faut pas négliger le risque dû aux erreurs d'arrondi d'obtenir un résultat faux. L'implantation avec une arithmétique entière exacte d'un quelconque algorithme de

[17]Voir le livre de Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford Science Publications, 1988.

stabilité pour traiter des polynômes dont les degrés par rapport à chacune des variables sont supérieur ou égaux à 5 n'est pas satisfaisante aujourd'hui : le temps nécessaire au test d'un polynôme est trop élevé. Dans un proche avenir, cela peut ne plus être vrai.

Les algorithmes (BB1) et (BB2) sont disponibles à l'URL

ftp ://ftp.metz.supelec.fr/pub/Supelec/software/auto/stability.tar.Z

ou depuis la page http ://www.metz.supelec.fr/metz/personnel/barret/logiciel.html.

Chapitre 3

Bancs de filtres adaptés, application au codage sans perte d'images

3.1 Introduction

Le travail de thèse de H. Bekkouche a porté sur le codage des images par bancs de filtres adaptés. Les structures de type *lifting scheme* que nous avons étudiées utilisent des filtres auto-régressifs à moyenne ajustée (ARMA) bi-dimensionnels dans les bancs de synthèse [30]. Pour être appliquées en compression sans perte, les transformations associées à ces structures sont rendues réversibles en ajoutant, comme pour les décompositions en ondelettes d'entiers en entiers, des arrondis à l'entier le plus proche juste après les filtres prédicteurs adaptés (voir la figure 3.1). L'introduction de ces non linéarités fait disparaître le problème du test de la stabilité des filtres de synthèse, car les opérations qui diffèrent entre le codeur et le décodeur ne portent que sur des variables entières et n'introduisent donc aucune erreur d'arrondi. En revanche, les opérations de filtrage sont réalisées, elles, avec une arithmétique à virgule flottante, mais étant parfaitement reproductibles elles sont identiques au codeur et au décodeur, n'introduisant pas de divergence entre codage et décodage.

Le codage des images par prédiction linéaire a été étudié au début (Chung & Kanesky, *IEEE Trans. Image Processing*, **IP-1**, 3, 1992 et sa bibliographie) et à la fin (Wu, Barthel & Zhang, *Proceedings of ICIP*, 1998) des années 90 sans être associé à des décompositions hiérarchiques et ne permettant donc pas de progressivité, ni en résolution, ni en qualité. Nous avons vu dans l'introduction de ce mémoire que la structure en *lifting scheme* de bancs de filtres à reconstruction parfaite introduit des filtres appelés malencontreusement prédicteurs, alors qu'ils ne correspondent qu'à des estimateurs linéaires : ils estiment un signal de sous-bande J_2 par la sortie d'un filtre appliqué à l'autre signal de sous-bande J_1. À la fin des années 90, Gerek et Çetin (*IEEE Trans. Image Processing*, **IP-9**, 10, 2000) ont étudié les performances en codage d'images des structures en *lifting scheme* dont les filtres prédicteurs s'adaptent à l'image selon l'algorithme du gradient stochastique. Comme nous l'avons déjà dit, cette approche n'exploite pas toute l'information disponible au décodeur pour prédire la valeur du signal J_2. Nous avons choisi avec H. Bekkouche et J. Oksman d'étudier les performances en codage d'images de structures du type *lifting scheme* où **toute** l'information disponible au décodeur est exploitée via des systèmes linéaires invariants par décalages (si l'on néglige les arrondis) dans l'étape dite de prédiction.

Autrement dit avec les notations du schéma de la figure 3.1, nous avons étudié l'impact du filtre $B(z_1, z_2)$ dans une étape de prédiction adaptée à l'image à coder : la valeur du pixel $J_2^0(m, n)$ est

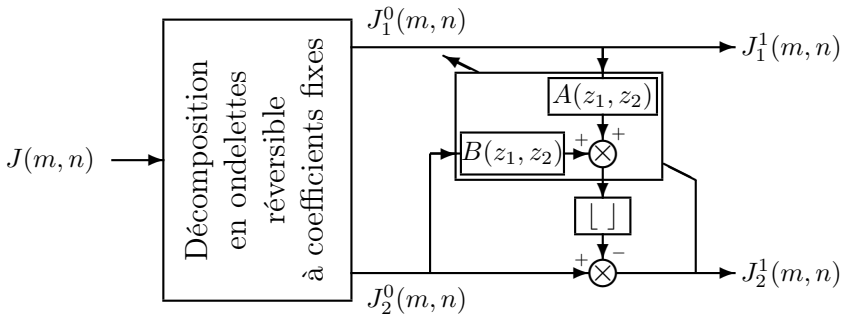


FIG. 3.1 – Structure de type *lifting scheme* étudiée avec H. Bekkouche, le symbole $\lfloor\ \rfloor$ désigne l'arrondi à l'entier le plus proche. Dans l'étape de prédiction, un filtre supplémentaire $B(z_1, z_2)$ a été ajouté à la structure en *lifting scheme* classique.

estimée par l'expression

$$\hat{J}_2^0(m,n) = \sum_{(h,k)\in\Delta_2} b_{h,k} J_2^0(m-h, n-k) + \sum_{(h,k)\in\Delta_1} a_{h,k} J_1^0(m-h, n-k),$$

où Δ_1 et Δ_2 sont deux voisinages bornés de $(0,0)$, le deuxième devant être semi-causal[1] et

$$J_2^1(m,n) = J_2^0(m,n) - \left\lfloor \hat{J}_2^0(m,n) + \frac{1}{2} \right\rfloor,$$

où pour un nombre réel x, $\lfloor x \rfloor$ désigne sa partie entière (i.e., le plus grand entier relatif inférieur ou égal à x).

Pour permettre un codage progressif en résolution, la décomposition en ondelettes à coefficients fixes qui a été utilisée dans nos tests est, soit une simple décomposition polyphase (*lazy wavelet*) pour les premiers niveaux de décomposition[2], soit la décomposition en ondelettes de Haar réversible[3] pour les niveaux de décomposition plus élevés (d'autres ondelettes réversibles pourraient être employées, comme la 5/3 de Daubechies qui est recommandée dans JPEG 2000).

Cette structure, intégrée dans une décomposition hiérarchique pyramidale, permet un codage progressif en résolution. Nous allons voir maintenant quel critère a été utilisé pour adapter la structure aux données et pourquoi.

3.2 Choix du critère pour une transformation réversible optimale en compression sans perte

Rappelons le schéma de principe d'un codeur par transformée et du décodeur associé (voir figure 3.2). L'objectif d'un tel dispositif est de réduire la taille du train binaire (encore appelé flot de bits), tout en maintenant une distorsion moyenne entre le signal initial et le signal décodé inférieure à un seuil donné. Dans le cas d'un codeur sans perte d'images, l'échantillonnage et la quantification sont réalisés par le système d'acquisition et ont lieu avant la transformation (nécessairement réversible), qui est encore suivie d'un codeur entropique. En général la quantification est scalaire et

[1] Une description plus détaillée de ces domaines peut être trouvée dans [29] ou dans [6].

[2] Les effets de repliement de spectre sont en général peu visibles et d'aucune gêne pour un aperçu rapide de l'image.

[3] Appelée également transformation S (*S-transform*).

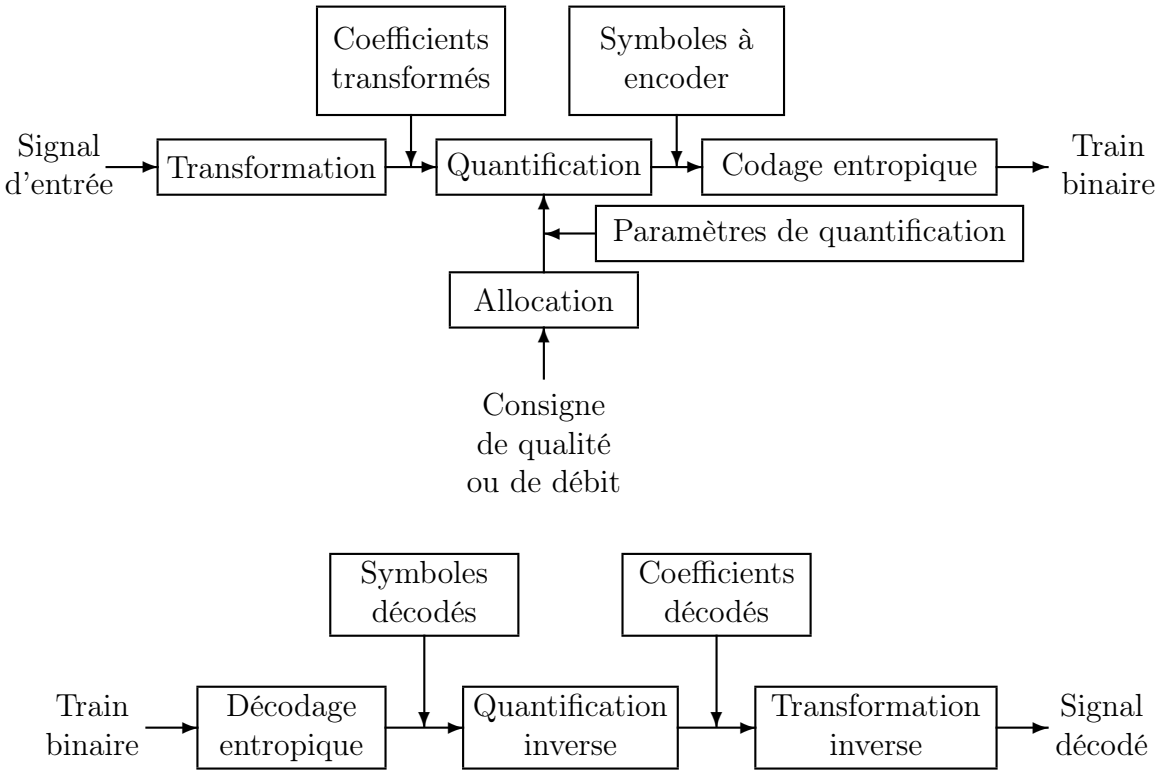


FIG. 3.2 – Codeur/décodeur par transformée.

uniforme avec un pas de quantification $q \ll 1$ qui vérifie les hypothèses d'une quantification dite *à haute résolution*. Dans le schéma ci-dessus, les symboles à encoder (à l'entrée du codeur entropique) peuvent être considérés comme issus d'une source aléatoire discrète (en général ils prennent des valeurs entières susceptibles de varier entre une valeur minimale et une valeur maximale connues). Pour ajuster en fonction de la source de symboles les distributions de probabilité du codeur entropique, il est nécessaire de la supposer ergodique et donc stationnaire (au moins localement). Il résulte du premier théorème de codage de Shannon que pour une source de symboles stationnaire (et fixée), le meilleur codeur entropique (i.e., minimisant la taille du flot de bits) donne comme taille moyenne du train binaire, exprimée en nombre de bits par symbole, le débit d'entropie de la source[4]. Par ailleurs, quand les symboles sont indépendants, le débit entropique coïncide avec l'entropie d'ordre 1 de la source (c'est-à-dire l'entropie $H(S_n)$). L'objectif de la transformation est de réduire le débit entropique de la source de symboles. Et, pour réduire la complexité du codeur entropique (il est plus simple pour un codeur d'approcher la valeur de l'entropie d'ordre 1 de la source plutôt que son débit entropique), la transformation est construite pour réduire à la fois la redondance entre symboles voisins et leur entropie d'ordre 1 (voire l'entropie conditionnelle sachant un petit voisinage causal du symbole courant).

[4]En notant $(S_n)_{n\in\mathbb{Z}}$ le processus aléatoire issu de la source de symboles, le débit d'entropie vaut

$$\lim_{T\to\infty} \frac{H(S_1,\ldots,S_T)}{T} = \lim_{T\to\infty} H(S_T \,|\, S_{T-1},\ldots,S_1),$$

où $H(S_1,\ldots,S_T)$ est l'entropie jointe des symboles $S_1,\ldots,S_T$ et $H(S_T \,|\, S_{T-1},\ldots,S_1)$ l'entropie conditionnelle de S_T sachant $S_{T-1},\ldots,S_1$.

Dans un système d'acquisition, l'image continue est échantillonnée (après avoir été filtrée pour vérifier les conditions du théorème d'échantillonnage de Shannon) puis quantifiée avec un quantificateur scalaire uniforme ayant le même pas de quantification pour tous les pixels de l'image (il existe en général un système de calibrage en amont qui ramène la plage des valeurs d'entrée de chaque capteur CCD entre 0 et 1). Une image échantillonnée peut être considérée comme une matrice $\mathbf{J}$ de dimension $M \times N$, l'élément situé à l'intersection de la ligne i et de la colonne j correspondant à la valeur du pixel localisé au point de coordonnées (i, j). En fixant un balayage de l'image, en général on part du coin en haut à gauche jusqu'au coin en bas à droite en parcourant les lignes une par une, cette matrice peut être identifiée à un vecteur de dimension $L = MN$, que l'on notera encore $\mathbf{J} = (J_1, \ldots, J_L)^T$. Chaque pixel subit une quantification scalaire uniforme de pas de quantification q, indépendant du pixel. Notons $Q_q[J_i]$ le résultat de la quantification scalaire uniforme de pas q appliquée à J_i et

$$Q_q[\mathbf{J}] = (Q_q[J_1], \ldots, Q_q[J_L])^T$$

l'image quantifiée. Les systèmes d'acquisition fournissent l'image $Q_q[\mathbf{J}]$ échantillonnée et quantifiée, l'image $\mathbf{J}$ est inaccessible à l'observation, mais il est très utile de l'introduire pour décrire le comportement de la transformation réversible. Quand sont omis les arrondis à l'entier le plus proche dans les décompositions par banc de filtres étudiées avec Hocine Bekkouche, ces dernières correspondent à des matrices triangulaires d'ordre L dont tous les éléments situés sur la diagonale principale valent 1. Notons $\mathbf{A}$ l'une de ces décompositions, $\mathbf{Y} = \mathbf{AJ} = (Y_1, \ldots, Y_L)^T$ l'image constituée des coefficients transformés sans quantification et

$$\mathrm{Rev}(\mathbf{A}, Q_q[\mathbf{J}]) = (Y_1^Q, \ldots, Y_L^Q)^T$$

le résultat de la transformation réversible[5] associée à $\mathbf{A}$ appliquée à l'image quantifiée $Q_q[\mathbf{J}]$. Remarquons que l'image $\mathbf{Y}$ est inaccessible à l'observation, tout comme $\mathbf{J}$. On peut montrer que sous l'hypothèse d'une quantification à haute résolution ($q \ll 1$), l'entropie d'ordre 1 du i-ème symbole issu de la transformation réversible est approximativement égale à l'entropie d'ordre 1 du i-ème coefficient transformé et quantifié, plus précisément :

$$H(Y_i^Q) = H(Q_q[Y_i]) + O(q^2) \qquad (\forall i \in [\![1\,;\, L]\!]).$$

La transformation réversible conserve l'entropie jointe :

$$H(Q_q[\mathbf{J}]) = H(\mathrm{Rev}(\mathbf{A}, Q_q[J]))$$

et il est tout aussi difficile de construire un codeur entropique qui donne une longueur moyenne du flot de bits égale à la valeur asymptotique $H(Q_q[\mathbf{J}])/L$ à partir de l'image quantifiée initiale $Q_q[\mathbf{J}]$ qu'à partir des coefficients transformés $\mathrm{Rev}(\mathbf{A}, Q_q[J])$. En revanche, quand on impose au codeur entropique d'approcher l'entropie d'ordre 1 à la place du débit entropique, la transformation $\mathbf{A}$ trouve son intérêt, car il n'y a pas conservation de la somme des entropies marginales (en général) :

$$\sum_{i=1}^{L} H(Q_q[J_i]) \neq \sum_{i=1}^{L} H(Y_i^Q).$$

Se pose alors le problème de trouver la transformation $\mathbf{A}$ dont la version réversible suivie d'un codeur entropique d'ordre 1 minimise la taille moyenne du train de bits. Autrement dit, chercher $\mathbf{A}$ telle

[5]En ajoutant les arrondis à l'entier le plus proche comme indiqué sur la figure 3.1.

que $\sum_{i=1}^{L} H(Y_i^Q)$ soit minimale. Or d'après ce qui précède, en introduisant l'information mutuelle $I(Y_1;\ldots;Y_L)$ entre les composantes de $\mathbf{Y}$, $h(Y_i)$ l'entropie (différentielle) de la variable aléatoire continue Y_i et $h(\mathbf{Y})$ l'entropie (différentielle) du vecteur $\mathbf{Y}$, il vient :

$$\begin{aligned}\sum_{i=1}^{L} H(Y_i^Q) &\simeq \sum_{i=1}^{L} H(Q_q[Y_i]) \simeq \sum_{i=1}^{L} h(Y_i) - L\log_2 q \\ &\simeq h(\mathbf{Y}) + I(Y_1;\ldots;Y_L) - L\log_2 q \simeq h(\mathbf{J}) + I(Y_1;\ldots;Y_L) - L\log_2 q\end{aligned}$$

car $H(Q_q[Y_i]) \simeq h(Y_i) - \log_2 q$ (formule de Bennett valable sous l'hypothèse d'une quantification à haute résolution) et $h(\mathbf{Y}) = h(\mathbf{J}) + \log_2 |\det \mathbf{A}| = h(\mathbf{J})$ (la matrice $\mathbf{A}$ est triangulaire avec des 1 sur la diagonale principale). Ainsi, la transformation $\mathbf{A}$ optimale est celle qui minimise l'information mutuelle $I(Y_1;\ldots;Y_L)$ entre coefficients transformés (sans arrondis à l'entier le plus proche, ni quantification) ou encore celle qui minimise la somme $\sum_{i=1}^{L} h(Y_i)$ des entropie différentielles marginales des coefficients transformés.

Dans le cas de données $\mathbf{J}$ gaussiennes, les coefficients transformés sont gaussiens de variance $\sigma_{Y_i}^2$ pour Y_i, et la transformation $\mathbf{A}$ optimale est celle qui minimise le produit des variances $\prod_{i=1}^{L} \sigma_{Y_i}^2$. En général les images issues de systèmes d'acquisition ne sont pas gaussiennes et la transformation $\mathbf{A}$ optimale (dans le schéma ci-dessus de compression sans perte par transformée) est celle qui minimise l'information mutuelle entre coefficients transformés, toutefois le critère à minimiser classiquement retenu en compression d'images est le produit des variances des coefficients transformés. Pour des données non gaussiennes, c'est une approximation qui donne déjà des résultats intéressants en pratique avec une complexité de calcul réduite par rapport au "vrai" critère.

Pour la structure de type *lifting scheme* étudiée avec Hocine Bekkouche et décrite à la figure 3.1, le critère qui a été retenu est l'erreur quadratique moyenne[6] : $\mathrm{E}[|J_2^0(m,n) - \hat{J}_2^0(m,n)|^2]$. Deux estimations de l'erreur quadratique moyenne ont été testées. L'une suppose l'image globalement stationnaire au sens large avec le critère des moindres carrés :

$$W_1 = \sum_{m=1}^{M_2} \sum_{n=1}^{N_2} \left| J_2^0(m,n) - \hat{J}_2^0(m,n) \right|^2$$

(où $M_2 \times N_2$ correspond à la dimension de l'image de sous-bande $\mathbf{J}_2$), et l'autre suppose l'image localement stationnaire au sens large avec le critère des moindres carrés adaptatifs :

$$W_2(m,n) = \sum_{i=1}^{m-1} \sum_{j=1}^{N_2} \left| J_2^0(i,j) - \hat{J}_2^0(i,j) \right|^2 \alpha^{N_2(m-i)+n-j} + \sum_{j=1}^{n} \left| J_2^0(m,j) - \hat{J}_2^0(m,j) \right|^2 \alpha^{n-j}.$$

Ces deux estimations de l'erreur quadratique moyenne ont donné lieu à deux méthodes d'adaptation : la méthode GAE (*Globally Adapted Estimation*) pour la première—décrite dans [30, 6]—et la méthode LAE (*Locally Adapted Estimation*) pour la deuxième—décrite dans [29, 6]. La raison qui nous a conduit à étudier ces différentes adaptations provient du fait qu'en général les modèles d'images ne sont pas appropriés à l'image complète. L'objectif était de construire des décompositions multirésolutions à partir de structures en *lifting scheme* capables de bonnes performances en compression sans perte d'images, en particulier pour les images satellitaires ou médicales.

[6] Comme cela est l'usage en compression d'images, c'est également le critère utilisé par Gerek & Çetin et par Wu, Barthel & Zhang.

3.3 Résultats obtenus et perspectives

Les méthodes présentées dans la thèse de H. Bekkouche (voir aussi [27] et sa bibliographie ainsi que [6]) permettent toutes les deux un codage sans perte et progressif en résolution. Dans la méthode LAE les coefficients du filtre prédicteur sont mis à jour à chaque pixel et ne sont pas transmis au décodeur (ce dernier les recalcule dans les mêmes conditions qu'au codeur). Le codeur et le décodeur ont la même complexité. En revanche, pour la méthode GAE, le même filtre prédicteur est appliqué à l'image entière et les coefficients des filtres doivent être transmis au décodeur. Dans ce cas, le décodeur a une complexité significativement plus faible que le codeur et tout à fait comparable à celle du décodeur de JPEG2000 [27, 6]. En compression sans perte, la structure en *lifting scheme* associée aux deux types d'adaptation ci-dessus donne des performances voisines de l'état de l'art [27]. En comparaison avec les autres codeurs permettant un codage progressif en résolution (i.e., `Jasper` une implantation de JPEG2000 et S+P de Said et Pearlman) les codeurs étudiés avec Hocine Bekkouche donnent un débit légèrement inférieur (quelques centièmes de bits par pixel) pour des images satellitaires [28] et des images de textures [6], mais au prix d'une plus grande complexité de codage. Nous avons vérifié également que, pour des images synthétiques vérifiant les hypothèses (images gaussiennes, localement ou globalement stationnaires au sens large) assurant l'optimalité du critère utilisé, la structure en *lifting scheme* généralisée (avec le filtre $B(z_1, z_2)$) donne des performances significativement meilleures que les autres bancs de filtres. Ce résultat très encourageant ne se retrouve pas pour les images naturelles ordinaires, montrant (une fois de plus) que l'hypothèse de données gaussiennes en codage d'images n'est pas très satisfaisante et ouvrant la voie, pour des travaux ultérieurs, à l'étude des structures en *lifting scheme* généralisé utilisant un autre critère d'adaptation : à savoir l'information mutuelle minimale.

Chapitre 4

Banc de filtres hybride et conversion analogique/numérique

Les systèmes de conversion analogique/numérique à base de bancs de filtres hybrides sont étudiés depuis la fin des années 1990 (Velasquez, Nguyen & Broadstone, *IEEE Trans. Signal Processing*, **SP-46**, 4, 1998 ; Löwenborg, Johansson & Wanhammar, *IEEE Int. Symp. Circuits Syst.*, Genève, Suisse, 2000). La motivation de cette recherche vient de la demande, rencontrée dans de nombreux domaines comme la communication sans fil, de taux de conversion de données toujours plus élevés. Malheureusement pour les bancs de filtres hybrides, leur grande sensibilité à de petites perturbations appliquées sur les valeurs de leurs composants analogiques, a réduit considérablement la faisabilité de tels systèmes (Petrescu & Oksman, soumis à *IEEE Trans. Circuits Systems*). Pour contourner ce problème, nous avons étudié de nouveaux bancs de filtres hybrides qui ont plus de degrés de liberté que les bancs classiques [24], ils correspondent à l'adaptation au cas hybride des bancs de filtres numériques non uniformes (i.e., dont le nombre K de sous-bandes diffère du facteur M de sur-échantillonnage). Nous espérons ainsi, que l'introduction de ces paramètres en excès permettra un meilleur contrôle de la sensibilité aux valeurs des composants analogiques. Avec J.-L. Collette, nous avons ajouté des canaux supplémentaires au banc de filtre hybride classique (décrit par exemple dans [25]) sans modifier le facteur de sur-échantillonnage (i.e., $K > M$), ou de façon équivalente nous avons réduit le facteur de sur-échantillonnage sans modifier le nombre de canaux. Nous avons alors montré comment les paramètres en excès pouvaient être utilisés pour minimiser la distorsion et le recouvrement (*aliasing*) sous la contrainte d'une amplification du bruit de quantification fixée [24]. Cela permet une amélioration significative des performances du banc de filtre hybride par rapport aux bancs classiques. La présentation des résultats est assez générale et peut s'étendre à d'autres contraintes, comme la sensibilité aux valeurs des composants analogiques.

Un dernier papier sur le sujet [19] présente l'étude théorique de l'erreur due d'une part à la reconstruction imparfaite du banc de filtre hybride non uniforme et d'autre part aux bruits de quantification des convertisseurs analogiques/numériques. Nous y proposons une nouvelle méthode de simulation qui évite la résolution numérique (et donc les erreurs de calculs associées) des équations différentielles des filtres analogiques et comparons les résultats des simulations aux prévisions théoriques. De plus, dans les études de performances des convertisseurs analogiques/numériques par bancs de filtres hybrides, il est d'usage de quantifier l'effet du bruit de quantification par sa puissance moyenne en sortie du banc, alors que le bruit de quantification est cyclo-stationnaire de période M. Nous donnons un exemple où la variance du bruit fluctue beaucoup sur une période de M échantillons et qui montre que la puissance moyenne peut être un mauvais indicateur de performance.

Chapitre 5

Analyse en composantes indépendantes et compression de données

5.1 Introduction

Grâce au financement de la thèse de Michel Narozny, dont j'ai été le directeur, nous avons poursuivi les travaux commencés avec Hocine Bekkouche et Jacques Oksman sur la compression d'images en nous focalisant sur le codage avec pertes pour des données non gaussiennes. La thèse de Michel Narozny traite de l'analyse en composantes indépendantes et de la compression de données. Les motivations de ce sujet ainsi que les publications liées à la thèse ont déjà été présentées dans l'introduction de ce mémoire.

Nous avons étudié le codage par transformée (décrit à la figure 3.2) et en particulier la description de la transformation optimale à hauts débits, **sans l'hypothèse de données gaussiennes**, du schéma de compression décrit à la figure 5.1. Les échantillons $x(n)$ du signal à coder sont regroupés par blocs de taille N sans recouvrement, formant ainsi un signal vectoriel $\mathbf{x}(m)$ à coder. Les échantillons du signal $\mathbf{x}(\cdot)$ sont des réalisations d'un vecteur aléatoire $\mathbf{X}$ de dimension N, auquel une transformation linéaire inversible $\mathbf{A}$ est appliquée pour obtenir le vecteur $\mathbf{Y}$. Chaque composante de ce vecteur est ensuite quantifiée et codée avec un codeur entropique, indépendamment des autres composantes, tout en assurant une allocation optimale de débits entre quantificateurs. Le décodeur commence par reconstruire une approximation $\hat{\mathbf{Y}}$ du vecteur transformé au moyen des N décodeurs entropiques et des N opérations de déquantification, avant de lui appliquer une transformation linéaire inversible $\mathbf{B}$ pour reconstruire une approximation $\hat{\mathbf{X}} = \mathbf{B}\hat{\mathbf{Y}}$ du vecteur $\mathbf{X}$.

Nous supposons qu'il y a exactement N quantificateurs (un par composante) suivis chacun d'un codeur entropique. Pour simplifier les expressions mathématiques, nous les supposons tous identiques, c'est-à-dire tous scalaires suivis d'un codeur entropique d'ordre V ou bien tous vectoriels de même dimension V suivis d'un codeur entropique d'ordre 1. Une allocation de débit optimale est effectuée entre les N quantificateurs pour réduire la distorsion (mesurée à partir d'un coût quadratique) entre le signal à coder $x(n)$ et le signal reconstruit $\hat{x}(n)$, sous la contrainte d'un débit total ne dépassant pas une valeur fixée à l'avance. En pratique, nous avons appliqué des quantificateurs scalaires et des codeurs entropiques d'ordre 1, mais l'étude théorique ne présente pas plus de difficultés dans le cas général. Enfin, nous imposons à la transformation linéaire du décodeur d'être l'inverse mathématique de celle du codeur : $\mathbf{B} = \mathbf{A}^{-1}$.

Avant de terminer ce paragraphe, introduisons les notations suivantes. Nous regroupons V échantillons consécutifs du vecteur $\mathbf{x}(\cdot)$ (resp. $\hat{\mathbf{x}}(\cdot)$) dans une matrice $\mathbf{x}^V(\ell) = [\mathbf{x}(\ell V), \mathbf{x}(\ell V+1), \ldots, \mathbf{x}(\ell V+V-1)]$ (resp. $\hat{\mathbf{x}}^V(\ell) = [\hat{\mathbf{x}}(\ell V), \hat{\mathbf{x}}(\ell V+1), \ldots, \hat{\mathbf{x}}(\ell V+V-1)]$) de dimensions $N \times V$ que nous supposons

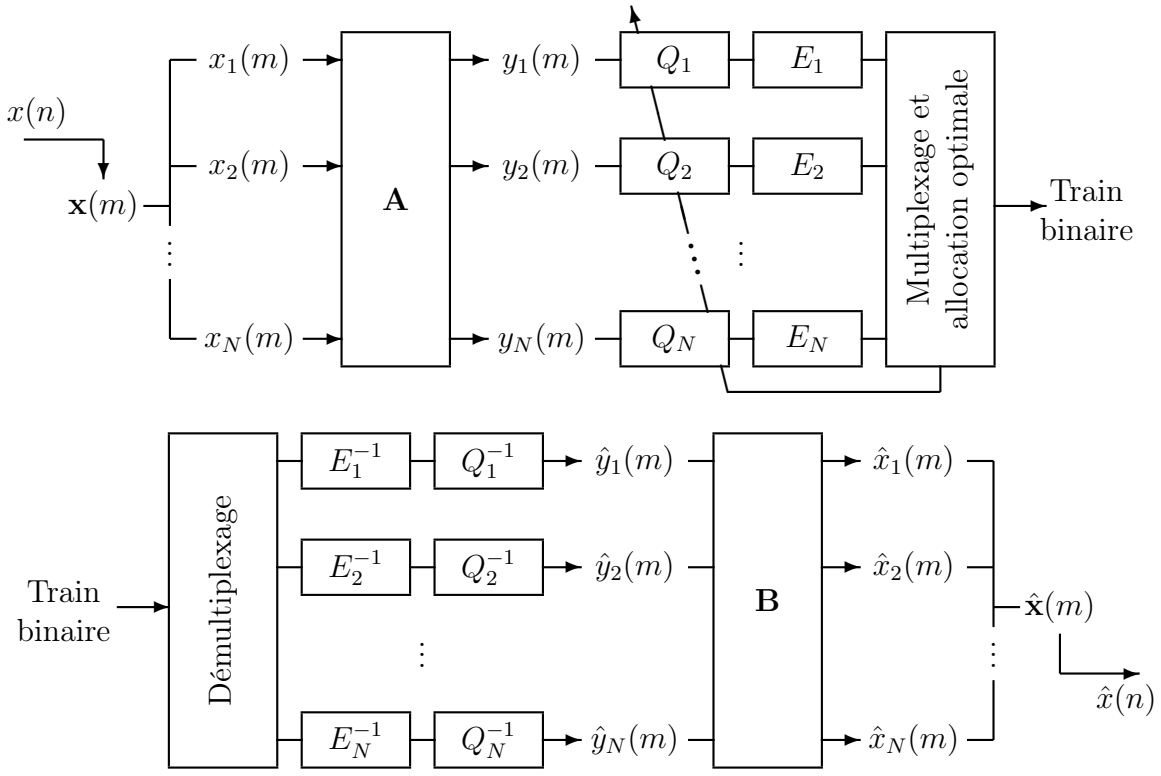


FIG. 5.1 – Codeur/décodeur par transformée. Pendant le codage le quantificateur Q_i suivi du codeur entropique E_i est appliqué à la $i^{\text{ème}}$ composante transformée y_i. Pendant le décodage le décodeur entropique E_i^{-1} (inverse mathématique de E_i) suivi de la déquantification Q_i^{-1} (qui n'est pas l'opération mathématique inverse de Q_i) est appliqué pour reconstruire une approximation $\hat{y}_i$ de la $i^{\text{ème}}$ composante transformée.

être une réalisation d'une matrice aléatoire $\mathbf{X}^V$ (resp. $\hat{\mathbf{X}}^V$), dont la $i^{\text{ème}}$ ligne est notée $\mathbf{X}_i^V$ (resp. $\hat{\mathbf{X}}_i^V$) et dont l'élément situé à l'intersection de la ligne i et de la colonne j est noté $X_{i,j}$ (resp. $\hat{X}_{i,j}$). Les mêmes notations sont appliquées aux coefficients transformés, c'est-à-dire aux signaux $\mathbf{y}(\cdot)$, $\hat{\mathbf{y}}(\cdot)$ et aux vecteurs aléatoires $\mathbf{Y}$ et $\hat{\mathbf{Y}}$. Enfin, pour une matrice $\mathbf{C}$, son élément situé à l'intersection de la ligne i et de la colonne j est noté $C_{i,j}$ ou $[C]_{i,j}$ et $\operatorname{tr}\mathbf{C}$, $\mathbf{C}^T$ désignent respectivement sa trace, sa transposée.

5.2 Gain de codage généralisé

5.2.1 Distorsion

Pour définir la distorsion, nous utilisons un coût quadratique avec un facteur de pondération $\alpha_i > 0$ par composante ($1 \leq i \leq N$) et nous cherchons à réduire la distorsion moyenne par échantillon

$$D(\mathbf{X}^V, \hat{\mathbf{X}}^V) = \frac{1}{N}\sum_{i=1}^{N} \alpha_i \left(\frac{1}{V}\sum_{j=1}^{V} \mathrm{E}\left[\left(X_{i,j} - \hat{X}_{i,j} \right)^2 \right] \right) \tag{5.1}$$

dans laquelle aucune région particulière du signal ou de l'image n'est favorisée (invariance par translation). En introduisant la matrice $\mathbf{F} = \operatorname{diag}(\sqrt{\alpha_1}, \ldots, \sqrt{\alpha_N})$ d'ordre N, où diag désigne la matrice diagonale constituée de son argument, la distorsion peut s'écrire

$$\begin{aligned} D(\mathbf{X}^V, \hat{\mathbf{X}}^V) &= \frac{1}{NV}\sum_{i=1}^{N}\sum_{j=1}^{V} \mathrm{E}\left\{ [\sqrt{\alpha_i}(X_{i,j} - \hat{X}_{i,j})]^2 \right\} \\ &= \frac{1}{NV} \operatorname{tr}\left[\mathbf{F}\, \mathrm{E}\big\{ (\mathbf{X}^V - \hat{\mathbf{X}}^V)(\mathbf{X}^V - \hat{\mathbf{X}}^V)^T \big\} \mathbf{F} \right] \\ &= \frac{1}{NV} \operatorname{tr}\left[\mathbf{F}\mathbf{B}\, \mathrm{E}\big\{ (\mathbf{Y}^V - \hat{\mathbf{Y}}^V)(\mathbf{Y}^V - \hat{\mathbf{Y}}^V)^T \big\} \mathbf{B}^T \mathbf{F} \right]. \end{aligned} \tag{5.2}$$

Nous supposons que pour chaque quantificateur le bruit de quantification est du deuxième ordre et centré et que les bruits de quantification des différents quantificateurs sont deux à deux décorrélés. Ces conditions sont généralement satisfaites pour une quantification à haute résolution (voir le paragraphe A.3). En notant

$$D_k(\mathbf{Y}_k^V, \hat{\mathbf{Y}}_k^V) = \frac{1}{V}\sum_{\ell=1}^{V} \mathrm{E}\left[\left(Y_{k,\ell} - \hat{Y}_{k,\ell} \right)^2 \right] = D_k,$$

la distorsion moyenne du $k^{\text{ème}}$ quantificateur, il résulte de l'hypothèse sur les bruits de quantification que

$$\frac{1}{V}\, \mathrm{E}\left\{ (\mathbf{Y}^V - \hat{\mathbf{Y}}^V)(\mathbf{Y}^V - \hat{\mathbf{Y}}^V)^T \right\} = \operatorname{diag}(D_1, \ldots, D_N) = \mathbf{D},$$

qui, avec l'équation (5.2), entraîne[1]

$$D(\mathbf{X}^V, \hat{\mathbf{X}}^V) = \frac{1}{N}\sum_{k=1}^{N} w_k D_k(\mathbf{Y}_k^V, \hat{\mathbf{Y}}_k^V) \quad \text{avec} \quad w_k = \sum_{i=1}^{N} \alpha_i U_{i,k}^2 = \left[\mathbf{B}^T\mathbf{F}^2\mathbf{B}\right]_{k,k}. \tag{5.3}$$

5.2.2 Allocation optimale de débits entre quantificateurs

Partant de l'expression (5.3) de la distorsion, la question qui nous intéresse ici est la suivante.

Disposant d'une distorsion totale maximale donnée $D(\mathbf{X}^V, \hat{\mathbf{X}}^V) = D$, comment la répartir entre les N composantes pour minimiser la somme des débits

$$R(D) = \frac{1}{N}\sum_{i=1}^{N} R_i(D_i)$$

(exprimée en bits par échantillons) des N quantificateurs ?

Il est bien connu (voir l'annexe A.1) que le débit minimal $R_i(D_i)$ par composante pour le schéma de compression de la figure 5.1 (sous l'hypothèse d'une faible distorsion D_i et d'une suite $(\mathbf{Y}_{i,k}^V)_{k\geq 1}$ indépendante et identiquement distribuée) vaut

$$R_i^{\text{Sh}}(D_i) = \inf\left\{ I(\mathbf{Y}_i^V\,;\,\hat{\mathbf{Y}}_i^V) \,:\, \frac{1}{V}\,\mathrm{E}\left[\|\mathbf{Y}_i^V - \hat{\mathbf{Y}}_i^V\|_2^2\right] \leq D_i \right\} \simeq \frac{h(\mathbf{Y}_i^V)}{V} - \frac{1}{2}\log_2[2\pi e D_i],$$

où $h(\mathbf{Y}_i^V)$ est l'entropie différentielle (voir l'annexe A.4) de $\mathbf{Y}_i^V$. Pour le schéma de compression de la figure 5.1 avec une quantification vectorielle en réseau à haute résolution, décrite par la matrice $\mathbf{U}_i$, suivie d'un codeur entropique d'ordre 1, le débit minimal vaut (voir l'annexe A.2)

$$R_i^{\text{Qv}}(D_i) = \frac{H(\hat{\mathbf{Y}}_i^V)}{V} \simeq \frac{h(\mathbf{Y}_i^V)}{V} - \frac{1}{2}\log_2\left[\frac{D_i}{G(\mathbf{U}_i)}\right],$$

où $H(\hat{\mathbf{Y}}_i^V)$ est l'entropie jointe de la composante quantifiée $\hat{\mathbf{Y}}_i^V$ et la grandeur $G(\mathbf{U}_i)$ est une constante qui ne dépend que de la forme du réseau du quantificateur (voir les annexes A.2 et A.6). Enfin, avec une quantification scalaire à haute résolution et un codeur entropique d'ordre V (voir les annexes A.3 et A.6), il vaut

$$R_i^{\text{Qs}}(D_i) = \frac{H(\hat{\mathbf{Y}}_i^V)}{V} \simeq \frac{h(\mathbf{Y}_i^V)}{V} - \frac{1}{2}\log_2[12 D_i].$$

Ainsi, pour les trois cas mentionnés ci-dessus le problème d'allocation optimale de débits est le même : il s'agit de trouver les distorsions moyennes D_i $(1 \leq i \leq N)$ qui minimisent la quantité

$$\Lambda(D_1, \ldots, D_N) = -\sum_{i=1}^{N} \log_2 D_i$$

[1]En effet, d'après les propriétés de la trace, on a

$$D(\mathbf{X}^V, \hat{\mathbf{X}}^V) = \frac{1}{N}\,\mathrm{tr}[\mathbf{FBDB}^T\mathbf{F}] = \frac{1}{N}\,\mathrm{tr}[\mathbf{DB}^T\mathbf{F}^2\mathbf{B}] = \sum_{i=1}^{N} \alpha_i\,\mathrm{tr}[\mathbf{DB}_i\,\mathbf{B}_i^T],$$

où $\mathbf{B}_i^T$ est la $i^{\text{ème}}$ ligne de $\mathbf{B}$. Ce qui entraîne $D(\mathbf{X}^V, \hat{\mathbf{X}}^V) = \frac{1}{N}\sum_{i=1}^{N} \alpha_i \mathbf{B}_i^T\mathbf{D}\mathbf{B}_i = \frac{1}{N}\sum_{k=1}^{N}\left(\sum_{i=1}^{N}\alpha_i U_{i,k}^2\right) D_k$.

sous la contrainte $\sum_{i=1}^{N} w_i D_i \leq ND$.

On montre [2] que l'allocation optimale entre quantificateurs est obtenue pour

$$D_{i,\text{opt}} = \frac{D}{w_i}. \tag{5.4}$$

Finalement, nous en déduisons (quel que soit le cas traité parmi les trois mentionnés ci-dessus) qu'avec une allocation optimale, le débit total minimal $R(D)$ associé à la distorsion totale D vérifie

$$R(D) \simeq \frac{1}{N}\sum_{i=1}^{N} \frac{h(\mathbf{Y}_i^V)}{V} - \frac{1}{2}\log_2 D + \frac{1}{2N}\sum_{i=1}^{N} \log_2 w_i + R_0 \tag{5.5}$$

où

$$R_0 = \begin{cases} -\frac{1}{2}\log_2[2\pi e] & \text{pour le débit asymptotique de Shannon} \\ \frac{1}{2N}\sum_{i=1}^{N} \log_2 G(\mathbf{U}_i) & \text{pour une quantification vectorielle à haute résolution} \\ -\frac{1}{2}\log_2 12 & \text{pour une quantification scalaire à haute résolution.} \end{cases}$$

Remarquons que l'approximation (5.5) est équivalente à la suivante

$$D(R) \simeq 2^{2(R_0-R)} \left(\prod_{i=1}^{N} w_i\right)^{\frac{1}{N}} \left(\prod_{i=1}^{N} 2^{2h(\mathbf{Y}_i^V)/V}\right)^{\frac{1}{N}}, \tag{5.6}$$

qui correspond à l'allocation optimale entre quantificateurs quand le débit est donné, c'est-à-dire à la distorsion $D(\mathbf{X}^V, \hat{\mathbf{X}}^V)$ minimale associée à un débit total inférieur à R.

Par ailleurs, $\mathbf{Y}^V$ est le résultat de la transformation linéaire $\mathbf{A}$ appliquée à chaque colonne de $\mathbf{X}^V$, ainsi nous avons

$$h(\mathbf{Y}^V) = h(\mathbf{X}^V) + V\log_2|\det \mathbf{A}|.$$

De plus, la somme des entropies différentielles des composantes vérifie $\sum_{i=1}^{N} h(\mathbf{Y}_i^V) = h(\mathbf{Y}^V) + I(\mathbf{Y}_1^V;\ldots;\mathbf{Y}_N^V)$, où $I(\mathbf{Y}_1^V;\ldots;\mathbf{Y}_N^V)$ est l'information mutuelle entre les composantes $\mathbf{Y}_1^V$, …, $\mathbf{Y}_N^V$ (voir l'annexe A.3). Donc, en ajoutant l'indice $\mathbf{A}$ au débit minimal pour indiquer qu'il dépend de la transformation, nous obtenons

$$R_{\mathbf{A}}(D) \simeq \frac{I(\mathbf{Y}_1^V;\ldots;\mathbf{Y}_N^V)}{NV} + \frac{\log_2|\det \mathbf{A}|}{N} + \frac{1}{2N}\sum_{i=1}^{N} \log_2 w_i + R_0(D), \tag{5.7}$$

où $R_0(D) = \frac{h(\mathbf{X}^V)}{NV} + R_0 - \frac{1}{2}\log_2 D$ est indépendant de $\mathbf{A}$.

5.2.3 Expression asymptotique de la réduction de débit

Nous proposons de définir la *réduction de débit* de la transformation $\mathbf{A}$ pour une distorsion D de la façon suivante :

$$\mathcal{R}(\mathbf{A}, D) = R_{\mathbf{A}_0}(D) - R_{\mathbf{A}}(D), \tag{5.8}$$

[2] En effet, minimiser $\Lambda(D_1,\ldots,D_N)$ revient à maximiser $\left(\prod_{i=1}^{N} D_i\right)^{1/N}$ ou encore à maximiser $\left(\prod_{i=1}^{N} w_i\right)^{1/N}\left(\prod_{i=1}^{N} D_i\right)^{1/N} = \left(\prod_{i=1}^{N} w_i D_i\right)^{1/N}$. La moyenne géométrique des $w_i D_i$ est majorée par leur moyenne arithmétique (elle même majorée par D d'après la contrainte) avec égalité si et seulement si tous les termes sont égaux, donc $w_i D_{i,\text{opt}} = D$ $(\forall i)$.

où $\mathbf{A}_0$ est une transformation étalon. En pratique, nous choisirons la transformation de Karhunen-Loève ou l'identité. Nous voyons qu'ainsi définie, la réduction de débit ne dépend pas du cas traité parmi les trois mentionnés ci-dessus et cela se généralise à une quantification vectorielle à haute résolution suivie d'un codeur entropique d'ordre quelconque.

Nous définissons également un *gain de codage généralisé* pour un débit total R :

$$\mathcal{G}(\mathbf{A}, R) = D_{\mathbf{A}_0}(R)/D_{\mathbf{A}}(R), \tag{5.9}$$

correspondant au facteur duquel la distorsion est réduite grâce à la transformation $\mathbf{A}$ par rapport à l'étalon $\mathbf{A}_0$. On a

$$\mathcal{R}(\mathbf{A}, D) = \frac{1}{2}\log_2 \mathcal{G}(\mathbf{A}, R).$$

Remarquons que pour $V = 1$, en introduisant la variance σ_i^2 de Y_i et la VA réduite $\tilde{Y}_i = \frac{Y_i}{\sigma_i}$ pour $1 \leq i \leq N$, il résulte de la relation $h(Y_i) = h(\tilde{Y}_i) + \log_2 \sigma_i$ que l'approximation (5.6) de la distorsion peut s'écrire

$$D(R) \simeq 2^{2(R_0-R)} \left(\prod_{i=1}^{N} c_i w_i\right)^{\frac{1}{N}} \left(\prod_{i=1}^{N} \sigma_i^2\right)^{\frac{1}{N}},$$

où $c_i = 2^{2h(\tilde{Y}_i)}$. Dans le cas où le signal à coder est gaussien, les coefficients c_i valent tous $2\pi e$, quelle que soit la transformation $\mathbf{A}$. De plus si la transformation $\mathbf{A}$ est orthogonale et si les coefficients de pondérations α_i de la fonction coût valent tous 1, alors les w_i valent aussi tous 1 et le gain de codage généralisé par rapport à l'identité coïncide avec le gain de codage $\left(\prod_{i=1}^{N} \mathrm{var}(X_i)\right)^{1/N} / \left(\prod_{i=1}^{N} \mathrm{var}(Y_i)\right)^{1/N}$ bien connu qui est maximisé par une transformée de Karhunen-Loève.

En utilisant la relation (5.7), on voit que maximiser la réduction de débit (ou de façon équivalente le gain de codage généralisé) revient à minimiser la quantité

$$\frac{I(\mathbf{Y}_1^V; \ldots; \mathbf{Y}_N^V)}{V} + \log_2 |\det \mathbf{A}| + \frac{1}{2}\sum_{i=1}^{N} \log_2 w_i,$$

de plus nous pouvons remarquer, d'après la relation (5.3), que

$$\log_2 |\det \mathbf{A}| + \frac{1}{2}\sum_{i=1}^{N} \log_2 w_i = \frac{1}{2}\log_2 \left[\frac{\det \mathrm{diag}(\mathbf{B}^T\mathbf{F}^2\mathbf{B})}{\det(\mathbf{B}^T\mathbf{F}^2\mathbf{B})}\right] + \log_2 \det \mathbf{F},$$

où appliquée à une matrice carrée la fonction diag désigne la matrice diagonale extraite de son argument. Finalement, comme nous l'a fait remarquer D. T. Pham, maximiser le gain de codage généralisé revient à minimiser le critère

$$\mathcal{C}(\mathbf{A}) = \frac{I(\mathbf{Y}_1^V; \ldots; \mathbf{Y}_N^V)}{V} + \frac{1}{2}\log_2 \left[\frac{\det \mathrm{diag}(\mathbf{A}^{-T}\mathbf{F}^2\mathbf{A}^{-1})}{\det(\mathbf{A}^{-T}\mathbf{F}^2\mathbf{A}^{-1})}\right], \tag{5.10}$$

qui se décompose en la somme de deux termes positifs ou nuls. En effet, l'information mutuelle $I(\mathbf{Y}_1^V; \ldots; \mathbf{Y}_N^V)$ est toujours positive et s'annule si et seulement si les vecteurs aléatoires $\mathbf{Y}_1^V, \ldots, \mathbf{Y}_N^V$ sont indépendants ; le deuxième terme est également toujours positif, d'après l'inégalité d'Hadamard, et s'annule si et seulement si la matrice $\mathbf{A}^{-T}\mathbf{F}^2\mathbf{A}^{-1}$ est diagonale, autrement dit si et seulement si les colonnes de la matrice $\mathbf{F}\mathbf{A}^{-1}$ sont deux à deux orthogonales.

Remarquons que le critère (5.10) est invariant par multiplication à gauche par une matrice diagonale inversible : en effet, si $\mathbf{D}$ est une matrice diagonale d'ordre N inversible, alors $\mathcal{C}(\mathbf{DA}) = \mathcal{C}(\mathbf{A})$ et cette propriété est intuitivement évidente, car l'adaptation du pas de quantification de chaque quantificateur permet de compenser l'amplification ou l'atténuation, due à la multiplication par la matrice $\mathbf{D}$, de la composante correspondante.

Remarquons également qu'en compression sans perte par transformée réversible (voir le chapitre 3), la transformation linéaire dont la version réversible sera optimale est celle qui minimise l'information mutuelle entre composantes du vecteur transformé, c'est-à-dire le premier terme du contraste ci-dessus. Le deuxième terme n'apparaît qu'en compression **avec** pertes quand la distorsion est l'erreur quadratique moyenne.

Dans le cas où $V = 1$, il existe un algorithme, `ICAinf` de Dinh-Tuan Pham, qui minimise le premier terme du critère ci-dessus ; et pour simplifier, nous avons étudié le cas où les coefficients α_i de l'équation (5.1) valent tous 1, i.e. quand $\mathbf{F}$ est la matrice identité. Nous obtenons ainsi le critère simplifié :

$$\mathcal{C}_s(\mathbf{A}) = I(Y_1; \ldots; Y_N) + \frac{1}{2} \log_2 \left[\frac{\det \operatorname{diag}(\mathbf{A}^{-T}\mathbf{A}^{-1})}{\det(\mathbf{A}^{-T}\mathbf{A}^{-1})} \right], \tag{5.11}$$

5.2.4 Lien avec l'analyse en composantes indépendantes

Un problème courant rencontré dans divers domaines comme l'analyse de données, le traitement du signal et la compression, est de trouver une représentation convenable de données multidimensionnelles. Pour des raisons de simplicité, une telle représentation est en général recherchée en appliquant une transformation linéaire aux données originales. Par exemple, on connaît depuis longtemps l'analyse en composantes principales, dont la matrice de changement de repère correspond à la transformation dite de Karhunen-Loève (TKL) dans les livres de codage, et qui donne une représentation où les composantes des données sont deux à deux décorrélées. Plus récemment, est apparue l'analyse en composantes indépendantes (ACI), pour laquelle la représentation recherchée est celle qui minimise l'information mutuelle entre les composantes des données. Parmi les méthodes existantes de séparation aveugle de sources ou d'analyse en composantes indépendantes, celles basées sur la minimisation de l'information mutuelle entre les sources estimées sont les plus appropriées au problème de codage (sous réserve de les modifier pour tenir compte du deuxième terme du critère $\mathcal{C}(\mathbf{A})$), car elles ne supposent pas que les observations sont obtenues à partir d'un mélange instantané de sources indépendantes éventuellement bruité. Grâce à l'aide précieuse de Dinh-Tuan Pham, avec Michel Narozny nous avons modifié son algorithme d'ACI, `ICAinf` (*IEEE Trans. Signal Processing*, **52**, 10, 2690–2700, 2004) qui retourne une matrice $\mathbf{A}$ minimisant l'information mutuelle $I(Y_1; \ldots; Y_N)$ entre composantes transformées via une méthode de descente de gradient relatif de type quasi-Newton, pour construire deux algorithmes : `OrthICA` et `GCGsup`, qui maximisent le gain de codage généralisé. Le premier impose à la matrice de séparation d'être orthogonale et l'autre est sans contrainte [23, 7].

Nous avons fait une analyse statistique des coefficients transformés (pour différentes transformations dont celles retournées par `OrthICA` et `GCGsup`) et une étude comparative à tous les débits, des transformations en codage avec pertes en s'assurant que l'allocation de bits entre quantificateurs est optimale.

Sur des signaux synthétiques nous avons obtenu des gains de codage importants (plus d'un bit par échantillon) par rapport à la TKL [23, 7]. Les premiers résultats obtenus sur des données réelles étaient moins encourageants, mais les derniers apparaissant dans le rapport de thèse de Michel Narozny et dans [7] le sont.

Toutefois, le codage par transformée d'images fixes, passant par une étape de découpage en blocs deux à deux disjoints de l'image, n'a plus la faveur des utilisateurs de codeurs d'images à cause des effets dits de blocs (apparition des frontières des blocs sur l'image reconstruite) à très bas débits. Le codage en sous-bandes (par exemple après une décomposition en ondelettes dyadique) est préféré, c'est ainsi que le récent standard JPEG2000 a supplanté l'ancien JPEG. Il serait plus intéressant, me semble-t-il, d'étudier des bancs de filtres associés à une décomposition dyadique pyramidale et optimaux (sous certaines hypothèses) en codage. Cela pourrait faire l'objet de futures recherches.

Chapitre 6

ACI appliquée au codage des images multicomposantes

6.1 Introduction

Grâce à l'Action Concertée Incitative Masse de Données, qui a accepté de financer le projet ACI^{2}M dont j'étais le coordinateur, nous avons pu poursuivre les travaux de Michel Narozny, bénéficiant ainsi du financement d'une thèse et des compétences de nos partenaires, spécialistes de haut niveau aussi bien en analyse en composantes indépendantes qu'en compression de données. La thèse d'Isidore Paul Akam Bita, dont j'ai été co-directeur avec Dinh-Tuan Pham, traite de l'application de l'analyse en composantes indépendantes (modifiée) au codage d'images multi- ou hyper-spectrales. Dans la continuité des travaux de Michel, Isidore Paul a commencé par associer des transformations à base d'analyse en composantes indépendantes pour réduire la redondance spectrale à des décompositions en ondelettes bi-dimensionnelles pour réduire la redondance spatiale. Les résultats obtenus ont donné lieu à trois présentations dans des ateliers ([40, 39, 37]), quatre articles dans des conférences internationales avec comité de lecture ([23, 38, 20, 17]) et la publication d'un papier long ([5]) dans la revue *Signal Processing*. Ces travaux se sont poursuivis après la thèse, conduisant à trois publications dans des conférences internationales avec actes et comité de lecture ([16, 14, 13]), à la publication d'un papier court ([4]) dans la revue *Signal Processing* et à la préparation d'un autre papier court à soumettre dans une revue [50].

La deuxième partie de la thèse de Paul a porté sur l'étude d'un modèle de mélange convolutif appliqué à la compression de données (en particulier d'images multicomposantes), les résultats obtenus ont fait l'objet d'une présentation dans un congrès avec actes et comité de lecture ([21]) et à la rédaction d'un article long [49] à soumettre pour publication dans une revue internationale. Les résultats prometteurs obtenus n'ont pas encore été exploités à leur juste valeur.

Les motivations du sujet de thèse d'Isidore Paul reposent d'une part sur une demande réelle de méthodes efficaces de compression d'images multispectrales et hyperspectrales, dont l'utilisation diverse et variée se répand dans de nombreux secteurs d'activité ; et d'autre part sur l'intérêt scientifique d'étudier de nouvelles méthodes de codage d'images dans le but de mieux comprendre les systèmes de codage rencontrés dans la Nature, par exemple dans le cerveau.

Les composantes d'une image multi-composante représentent généralement la même scène avec différentes vues dépendant de la longueur d'onde, à d'éventuels légers décalages près (correspondant au phénomène de *déregistration*). Il existe ainsi une forte dépendance entre les composantes, dite redondance spectrale et dans chaque composante, comme toute image, une forte redondance spatiale entre pixels voisins. Ces dernières décennies, différentes solutions ont été proposées pour réduire les

redondances, spatiale et spectrale, en compression d'images multicomposantes. Une solution assez répandue consiste à appliquer deux transformations distinctes, une par type de redondance. Par exemple, chaque composante subit une décomposition en ondelettes 2D pour diminuer la redondance spatiale et la spectrale est réduite par une application linéaire entre composantes. Une autre solution consiste à appliquer une décomposition en ondelettes 3D. Ces deux approches sont compatibles avec la partie 2 du standard JPEG2000.

Nous nous sommes focalisés dans un premier temps sur la première approche en étudiant trois schémas de compression (dont l'un est compatible JPEG2000), pour lesquels nous avons explicité un critère que doit minimiser une transformation optimale et donné deux algorithmes qui retournent une matrice minimisant le critère, l'un avec la contrainte d'une matrice orthogonale et l'autre sans autre contrainte que d'être inversible.

6.2 Description de trois schémas de compression

6.2.1 Conventions et notations

Nous considérons une image multicomposante $\mathbf{X}$ avec N composantes $\mathbf{X}_1, \ldots, \mathbf{X}_N$. Chaque composante $\mathbf{X}_i$ est une image 2D de dimension $N_r \times N_c$. Pour simplifier les expressions mathématiques, nous supposons que chaque composante est écrite sous la forme d'un vecteur ligne obtenu en balayant tous les pixels d'une composante dans un ordre arbitraire mais fixé et commun à toutes les composantes (par exemple de gauche à droite dans chaque ligne et ligne après ligne en commençant par celle du haut). Ainsi, $\mathbf{X}$ est une matrice de dimension $N \times L$ avec $L = N_c N_r$. Dans la suite, suivant le contexte, nous interpréterons $\mathbf{X}_i$ comme une image 2D ou une matrice ligne de dimension L.

L'opérateur linéaire, qui transforme une matrice en un vecteur colonne en empilant ses colonnes les unes au-dessus des autres de la première (qui se retrouve en haut) à la dernière (en bas), est noté vec. Soit $P > 0$ un entier naturel et KP un multiple de P, l'opérateur linéaire qui à la matrice $\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_K]$ de dimension $N \times KP$ (où chaque bloc $\mathbf{B}_i$ est de dimension $N \times P$) associe la matrice $[\mathrm{vec}(\mathbf{B}_1), \ldots, \mathrm{vec}(\mathbf{B}_K)]$ de dimension $NP \times K$ est noté perm et perm^{-1} désigne son inverse mathématique. Pour une matrice carrée $\mathbf{M}$, les expressions $\det \mathbf{M}$ et $\mathrm{diag}(\mathbf{M})$ désignent respectivement son déterminant et la matrice diagonale ayant les mêmes éléments diagonaux.

Dans les schémas de compression ci-dessous, les décompositions en ondelettes (DO) 2D ont toutes des coefficients fixes indépendants des données, contrairement aux transformations spectrales qui s'adaptent aux données. La même DO est appliquée à chaque composante, elle est représentée par la matrice $\mathbf{W}$ de dimension $L \times L$. Dans tous nos tests nous avons utilisé la DO dite 9/7 de Daubechies.

6.2.2 Schéma séparable

Le schéma séparable est compatible avec la partie 2 du standard JPEG2000, en voici la description.

– Codage : chaque composante $\mathbf{X}_i$ subit la même DO 2D, correspondant pour l'image complète au produit $\mathbf{X}\mathbf{W}^T$, et une transformation linéaire $\mathbf{A}$ est appliquée entre les composantes pour réduire la redondance spectrale. Les coefficients transformés sont ainsi les éléments de la matrice $\mathbf{Y} = \mathbf{A}\mathbf{X}\mathbf{W}^T$. Puis les coefficients transformés sont quantifiés et codés sans perte par un codeur entropique.

– Décodage : soit $\widehat{\mathbf{Y}}$ la matrice de même dimension que $\mathbf{Y}$ constituée des coefficients transformés quantifiés (on dit parfois déquantifiés). Les transformations inverses du codeur sont alors appliquées à $\widehat{\mathbf{Y}}$ pour reconstruire une approximation $\widehat{\mathbf{X}} = \mathbf{A}^{-1}\widehat{\mathbf{Y}}\mathbf{W}^{-T}$ de l'image originale.

Remarquons que l'ordre dans lequel les transformations $\mathbf{W}$ et $\mathbf{A}$ sont appliquées ne change pas le résultat : $\mathbf{Y} = \mathbf{A}(\mathbf{X}\mathbf{W}^T) = (\mathbf{A}\mathbf{X})\mathbf{W}^T = \mathbf{A}\mathbf{X}\mathbf{W}^T$, c'est pourquoi ce schéma de compression est dit *séparable*.

6.2.3 Deux variantes non séparables du schéma séparable

Il est bien connu que les statistiques des coefficients d'ondelettes dépendent de la sous-bande où ils se situent. Dans le schéma séparable, si la transformation spectrale s'adapte aux données, elle ne peut pas tenir compte pleinement de cette différence de statistiques, car la même transformation est appliquée à toutes les sous-bandes. Les schémas suivants n'a pas ce défaut, mais ils ne sont plus compatibles avec la partie 2 du standard JPEG2000.

Le schéma en sous-bandes

Dans le schéma en sous-bandes, l'image à coder doit subir d'abord les décompositions en ondelettes 2D, avant qu'une transformation spectrale spécifique à chaque sous-bande soit appliquée. Cela justifie le terme de schéma en *sous-bandes*, dont voici la description.

- Codage : en premier lieu, la même DO est appliquée à chaque composante, ainsi les coefficients d'ondelettes sont les éléments de la matrice $\mathbf{X}\mathbf{W}^T$. Deuxièmement, pour chaque composante les coefficients d'ondelettes sont regroupés par sous-bande suivant un balayage pré-défini commun à toutes les composantes. Ce re-ordonnancement correspond à la multiplication à droite de $\mathbf{X}\mathbf{W}^T$ par une matrice de permutation $\mathbf{P}^T$. Nous pouvons supposer sans perte de généralité que cette permutation est l'identité, quite à remplacer $\mathbf{W}$ par $\mathbf{P}\mathbf{W}$. Ce regroupement peut s'écrire $\mathbf{X}\mathbf{W}^T = [(\mathbf{X}\mathbf{W}^T)^{(1)} \cdots (\mathbf{X}\mathbf{W}^T)^{(M)}]$, où M est le nombre de sous-bandes par composante. Troisièmement, une transformation linéaire spécifique à chaque sous-bande est appliquée ($\mathbf{A}^{(m)}$ pour la sous-bande m). Les coefficients transformés sont alors les éléments de la matrice $\mathbf{Y} = [\mathbf{A}^{(1)}(\mathbf{X}\mathbf{W}^T)^{(1)} \cdots \mathbf{A}^{(M)}(\mathbf{X}\mathbf{W}^T)^{(M)}]$. Finalement, les coefficients transformés sont quantifiés et codés sans perte par un codeur entropique.
- Décodage : soit $\widehat{\mathbf{Y}} = [\widehat{\mathbf{Y}}^{(1)}, \ldots, \widehat{\mathbf{Y}}^{(M)}]$ la matrice des coefficients transformés quantifiés (on dit parfois déquantifiés), regroupés par sous-bande. Les transformations inverses de celles du codeur sont appliquées donnant une approximation

$$\widehat{\mathbf{X}} = [{\mathbf{A}^{(1)}}^{-1}\widehat{\mathbf{Y}}^{(1)}, \ldots, {\mathbf{A}^{(M)}}^{-1}\widehat{\mathbf{Y}}^{(M)}]\mathbf{W}^{-T}$$

  de l'image originale.

Le schéma mixte en sous-bandes

Il est bien connu que des redondances spatiales persistent après une DO 2D. Le codeur entropique EBCOT de JPEG2000 exploite cette redondance. Dans le schéma en sous-bande, les transformations spectrales diminuent la redondance spectrale mais ne modifient pas la redondance spatiale résiduelle. Le schéma mixte en sous-bande rectifie cet inconvénient. La différence avec le schéma en sous-bandes réside dans le fait que les transformations dites spectrales sont utilisées pour réduire la redondance spatiale. Cela est réalisé en ajoutant après la deuxième étape ci-dessus une décomposition polyphase de chaque composante en P composantes de plus faible dimension. Pour pouvoir appliquer l'opérateur perm (qui réalise la décomposition polyphase) à chaque sous-bande de chaque composante, nous supposons que le nombre de coefficients d'ondelettes par sous-bande et par composante est un multiple de P et donc L est divisible par P ; posons $L = KP$. Le nom de ce schéma provient d'une réduction

de redondance *mixte*, spatiale et spectrale, visée par les transformations "spectrales", en voici la description.

– Codage : en premier lieu la même DO est appliquée à chaque composante, les coefficients d'ondelettes constituant les éléments de la matrice $\mathbf{X}\mathbf{W}^T$. En deuxième, pour chaque composante les coefficients d'ondelettes sont regroupés par sous-bande suivant un balayage pré-défini commun à toutes les composantes ; le balayage dépend également du facteur de sous-échantillonnage P : il doit être tel que P coefficients d'ondelettes voisins (par exemple dans un bloc 2×2 quand $P = 4$) se retrouvent comme P coefficients d'ondelettes consécutifs après le ré-ordonnancement. Ce dernier correspond à la multiplication à droite de $\mathbf{X}\mathbf{W}^T$ par une matrice de permutation que l'on peut supposer être l'identité, sans perte en généralité (comme pour le schéma en sous-bandes). En troisième, l'opérateur perm est appliqué à chaque sous-bande :

$$\left[(\mathbf{X}\mathbf{W}^T)^{(1)} \dots (\mathbf{X}\mathbf{W}^T)^{(M)}\right] \longmapsto \left[\mathrm{perm}\left\{(\mathbf{X}\mathbf{W}^T)^{(1)}\right\} \dots \mathrm{perm}\left\{(\mathbf{X}\mathbf{W}^T)^{(M)}\right\}\right].$$

Puis, une transformation spectrale spécifique à chaque sous-bande est appliquée ($\mathbf{A}^{(m)}$, de dimension $NP \times NP$, pour la sous-bande m). Les coefficients transformés sont les éléments de la matrice

$$\mathbf{Y} = \left[\mathbf{A}^{(1)}\,\mathrm{perm}\left\{(\mathbf{X}\mathbf{W}^T)^{(1)}\right\} \dots \mathbf{A}^{(M)}\,\mathrm{perm}\left\{(\mathbf{X}\mathbf{W}^T)^{(M)}\right\}\right]$$

de dimension $NP \times K$. Enfin, les coefficients transformés sont quantifiés et codés sans perte par un codeur entropique.

– Décodage : soit $\widehat{\mathbf{Y}} = [\widehat{\mathbf{Y}}(1) \dots \widehat{\mathbf{Y}}^{(M)}]$ la matrice constituée des coefficients transformés quantifiés et regroupés par sous-bande. Les transformations inverses sont appliquées à $\widehat{\mathbf{Y}}$ pour reconstruire une approximation $\widehat{\mathbf{X}}$ de l'image originale :

$$\widehat{\mathbf{X}} = \left[\mathrm{perm}^{-1}\left\{\mathbf{A}^{(1)^{-1}}\widehat{\mathbf{Y}}^{(1)}\right\} \dots \mathrm{perm}^{-1}\left\{\mathbf{A}^{(M)^{-1}}\widehat{\mathbf{Y}}^{(M)}\right\}\right] \mathbf{W}^{-T}.$$

Nous noterons K_m le nombre de colonnes de la matrice $\mathbf{Y}$ associées à la sous-bande m.

Remarquons que le schéma en sous-bande est un cas particulier du schéma mixte en sous-bandes (quand $P = 1$) et que le schéma séparable est un cas particulier du schéma en sous-bandes (quand toutes les transformations spectrales sont identiques).

6.3 Expressions de la distorsion

Pour pouvoir appliquer le même type de raisonnement qu'au chapitre précédent, il est important de disposer d'une formule qui exprime la distorsion (erreur quadratique moyenne) finale entre l'image reconstruite et l'image originale en fonction des distorsions de chaque quantificateur. Pour cela, introduisons les hypothèses suivantes :

$\mathcal{H}_1$: *les composantes du bruit de quantification* $\mathbf{Y} - \widehat{\mathbf{Y}}$ *sont centrées et deux à deux décorrélées.*

$\mathcal{H}_2$: *la matrice* $\mathbf{W}^{-T}\mathbf{W}^{-1}$ *est diagonale.*

$\mathcal{H}_3$: *la quantité* $\|\mathbf{W}^{-1}\mathbf{e}_{(k-1)P+u}\|^2$*, où* $1 \le k \le K$*,* $1 \le u \le P$ *et* $\mathbf{e}_n$ $(1 \le n \le L)$ *est le* n*-ème vecteur de la base canonique de* $\mathbb{R}^L$*, ne dépend pas de la position spatiale dans la sous-bande.*

La condition $\mathcal{H}_3$ signifie que pour k, k', u, u' ($1 \le k, k' \le K$ et $1 \le u, u' \le P$) quelconques, si $(k-1)P+u$ et $(k'-1)P+u'$ sont des indices de colonnes de $\mathbf{Y}$ associés à la même sous-bande, alors $\|\mathbf{W}^{-1}\mathbf{e}_{(k-1)P+u}\|^2 = \|\mathbf{W}^{-1}\mathbf{e}_{(k'-1)P+u'}\|^2$.

Remarquons que la condition $\mathcal{H}_1$ est approximativement satisfaite (avec une bonne précision) sous les hypothèses d'une quantification à haute résolution (voir les annexes A.2 et A.3). La condition $\mathcal{H}_2$

est satisfaite par toute DO orthogonale, mais semble très restrictive, car seule l'ondelette de Haar est orthogonale, symétrique et à support compact. Toutefois, l'ondelette 9/7 de Daubechies, que nous avons utilisée dans tous nos tests, est presqu'orthogonale : précisément la matrice $\mathbf{W}^{-T}\mathbf{W}^{-1}$ est presque diagonale, (ses éléments non diagonaux sont relativement petits par rapport aux diagonaux). La condition $\mathcal{H}_3$ est satisfaite par des ondelettes dyadiques ayant des filtres de synthèse à réponse impulsionnelle finie, quand les effets de bord sont négligés (pour plus de détails, lire par exemple l'article de Woods et Naven, "A Filter Based Bit Allocation Scheme for Subband Compression of HDTV", *IEEE Trans. on Image Processing*, **1** (3), pp. 436–440, 1992).

Nous avons alors montré les théorèmes suivants (corollaires 1, 2 et 3 de [5]).

Théorème 10 (Schéma mixte en sous-bandes) *Avec un quantificateur par ligne de la matrice* $\mathbf{Y}$ *et par sous-bande (c'est-à-dire par composante, quand il y en a* NP *après la décomposition polyphase d'un facteur* P*, et par sous-bande) et*

1. *sous les hypothèses* $\mathcal{H}_1$*,* $\mathcal{H}_2$ *et* $\mathcal{H}_3$ *la distorsion finale* $D = \frac{1}{NL}\sum_{i=1}^{N}\sum_{n=1}^{L}(\mathbf{X}_i(n) - \widehat{\mathbf{X}}_i(n))^2$ *vérifie*

$$D = \frac{1}{NP}\sum_{j=1}^{N}\sum_{\ell=1}^{P}\sum_{m=1}^{M}\pi_m\omega_m w_{j,\ell}^{(m)} D_{j,\ell}^{(m)}, \tag{6.1}$$

   avec $\pi_m = \frac{K_m}{K} = \frac{PK_m}{L}$ *la proportion des coefficients d'ondelettes dans la sous-bande* m*,* $\omega_m = \frac{1}{K_m}\sum_k \|\mathbf{W}^{-1}\mathbf{e}_{kP}\|^2$ *(où dans la somme la plage de variation de* k *consiste en les* K_m *indices de colonnes de* $\mathbf{Y}$ *associés à la sous-bande* m*),* $w_{j,\ell}^{(m)} = \|\mathbf{A}^{(m)^{-1}}\mathbf{e}_{(\ell-1)N+j}\|^2$ *et* $D_{j,\ell}^{(m)}$ *la distorsion moyenne du* $\ell^{ème}$ *élément* $(1 \le \ell \le P)$ *d'un bloc localisé dans la sous-bande* m *de la* $j^{ème}$ *composante.*

2. *si la DO* $\mathbf{W}$ *et les transformations spectrales* $\mathbf{A}^{(m)}$ *sont toutes orthogonales, alors la relation (6.1) reste vraie sans l'hypothèse* $\mathcal{H}_1$ *et dans ce cas nous avons* $\omega_m = w_{j,\ell}^{(m)} = 1$ *pour tous* j*,* ℓ *et* m*.*

Théorème 11 (Schéma en sous-bandes) *Avec un quantificateur par sous-bande et par composante,*

1. *sous l'hypothèse* $\mathcal{H}_1$*, la distorsion finale du schéma en sous-bandes vérifie*

$$D = \frac{1}{N}\sum_{m=1}^{M}\pi_m\omega_m\left[\sum_{i=1}^{N} w_i^{(m)} D_i^{(m)}\right], \tag{6.2}$$

   où π_m *est la proportion des coefficients d'ondelettes dans la sous-bande* m*,* $\omega_m = \frac{1}{K_m}\sum_k \|\mathbf{W}^{-1}\mathbf{e}_k\|^2$ *(dans la somme, la plage de variation de* k *consiste en les* K_m *indices de colonnes de* $\mathbf{Y}$ *associés à la sous-bande* m*),* $w_i^{(m)} = \|\mathbf{A}^{(m)^{-1}}\mathbf{e}_i\|^2$ *et* $D_i^{(m)}$ *est la distorsion du quantificateur de la sous-bande* m *de la* $i^{ème}$ *composante.*

2. *si la DO* $\mathbf{W}$ *et les transformations spectrales* $\mathbf{A}^{(m)}$ *sont toutes orthogonales, alors la relation (6.2) reste vraie sans l'hypothèse* $\mathcal{H}_1$ *et dans ce cas nous avons* $\omega_m = w_i^{(m)} = 1$ *pour tous* i *et* m*.*

Théorème 12 (Schéma séparable) *Avec un quantificateur par sous-bande et par composante,*

1. *sous l'hypothèse* $\mathcal{H}_1$*, la distorsion finale du schéma séparable vérifie :*

$$D = \frac{1}{N}\sum_{m=1}^{M}\pi_m\omega_m\left[\sum_{i=1}^{N} w_i D_i^{(m)}\right], \tag{6.3}$$

où π_m est la proportion des coefficients d'ondelettes dans la sous-bande m, $\omega_m = \frac{1}{K_m}\sum_k \|\mathbf{W}^{-1}\mathbf{e}_k\|^2$ (dans la somme, la plage de variation de k consiste en les indices des colonnes de $\mathbf{Y}$ associés à la sous-bande m), $w_i = \|\mathbf{A}^{-1}\mathbf{e}_i\|^2$ et $D_i^{(m)}$ est la distorsion du quantificateur de la sous-bande m de la $i^{ème}$ composante.

2. *Si la DO $\mathbf{W}$ et la transformation spectrale $\mathbf{A}$ sont orthogonales, alors la relation (6.3) reste vraie sans l'hypothèse $\mathcal{H}_1$ et dans ce cas nous avons $\omega_m = w_i = 1$ pour tous i et m.*

Pour chaque schéma de compression, nous avons recherché la transformation spectrale optimale, c'est-à-dire celle qui minimise le débit sous la contrainte d'une distorsion finale maximale donnée, qui s'adapte aux données quand la DO est fixée. Nous avons commencé par expliciter un critère que doit minimiser toute transformation optimale.

6.4 Critères minimisés par les transformations optimales

En appliquant le même raisonnement qu'au chapitre précédent, qui rappelons-le est basé sur l'approximation de Bennett, nous avons démontré les trois théorèmes suivants (théorème 3, corollaire 4 et théorème 4 de [5]). La notation $h(Y_{j,\ell}^{(m)})$ désigne l'entropie différentielle des coefficients transformés $Y_{j,\ell}^{(m)}$ localisés en position ℓ ($1 \leq \ell \leq P$) des blocs de dimension P de la m-ème sous-bande ($1 \leq m \leq M$) de la i-ème composante ($1 \leq i \leq N$).

Théorème 13 (Schéma mixte en sous-bandes) *Quand la DO 2D est fixe, sous les hypothèses d'une quantification à haute résolution, $\mathcal{H}_2$ et $\mathcal{H}_3$, la transformation spectrale optimale $\mathbf{A}^{(m)}$ associée à la sous-bande m est une matrice $NP \times NP$ qui minimise le critère*

$$C_1(\mathbf{A}^{(m)}) = \sum_{j=1}^{N}\sum_{\ell=1}^{P} h(Y_{j,\ell}^{(m)}) + \frac{1}{2}\log_2 \det \operatorname{diag}\left[\mathbf{A}^{(m)-T}\mathbf{A}^{(m)-1}\right]. \tag{6.4}$$

Corollaire 14 (Schéma en sous-bandes) *Quand la DO 2D est fixe, sous la seule hypothèse d'une quantification à haute résolution, la transformation spectrale optimale $\mathbf{A}^{(m)}$ associée à la sous-bande m est une matrice $N \times N$ qui minimise le critère*

$$C_1(\mathbf{A}^{(m)}) = \sum_{j=1}^{N} h(Y_j^{(m)}) + \frac{1}{2}\log_2 \det \operatorname{diag}\left[\mathbf{A}^{(m)-T}\mathbf{A}^{(m)-1}\right].$$

Remarquons que pour les deux schémas non séparables, le critère $C_1(\mathbf{A}^{(m)})$, où l'exposant $^{(m)}$ a été omis pour alléger les notations, se met sous la forme

$$\begin{aligned} C_1(\mathbf{A}) &= \sum_{i=1}^{N} h(Y_i) - \log_2 \det \mathbf{A} + \frac{1}{2}\log_2 \frac{\det \operatorname{diag}(\mathbf{A}^{-T}\mathbf{A}^{-1})}{\det(\mathbf{A}^{-T}\mathbf{A}^{-1})} \\ &= I(Y_1;\ldots;Y_n) + h(\mathbf{X}\mathbf{W}^T) + \frac{1}{2}\log_2 \frac{\det \operatorname{diag}(\mathbf{A}^{-T}\mathbf{A}^{-1})}{\det(\mathbf{A}^{-T}\mathbf{A}^{-1})}, \end{aligned}$$

on retrouve ainsi, au terme $h(\mathbf{X}\mathbf{W}^T)$ (indépendant de $\mathbf{A}$) près, le critère du chapitre précédent que minimise les transformations optimales en compression d'images fixes. Les algorithmes `GCGsup` et `OrthICA` introduits au chapitre précédent permettent donc de calculer les transformations optimales des deux schémas non séparables.

Théorème 15 (Schema séparable) *Quand la DO 2D est fixe, sous les hypothèses d'une quantification à haute résolution, une transformation spectrale optimale* $\mathbf{A}$ *est une matrice de dimension* $N \times N$ *qui minimise le critère :*

$$C_2(\mathbf{A}) = \sum_{j=1}^{N} \sum_{m=1}^{M} \pi_m h(Y_j^{(m)}) + \frac{1}{2} \log_2 \det \operatorname{diag}(\mathbf{A}^{-T}\mathbf{A}^{-1}). \tag{6.5}$$

Puisque $\sum_{m=1}^{M} \pi_m = 1$, nous pouvons remarquer que le critère C_2 se met sous la forme

$$\begin{aligned} C_2(\mathbf{A}) &= \sum_{m=1}^{M} \pi_m \left[\sum_{i=1}^{N} h(Y_i^{(m)}) - \log_2 |\det \mathbf{A}| \right] + \frac{1}{2} \log_2 \left[\frac{\det \operatorname{diag}\left(\mathbf{A}^{-T}\mathbf{A}^{-1}\right)}{\det(\mathbf{A}^{-T}\mathbf{A}^{-1})} \right] \\ &= \sum_{i=1}^{M} \pi_m \left[I(Y_1^{(m)}; \ldots; Y_N^{(m)}) + h\big((\mathbf{X}\mathbf{W}^T)^{(m)}\big) \right] + \frac{1}{2} \log_2 \left[\frac{\det \operatorname{diag}\left(\mathbf{A}^{-T}\mathbf{A}^{-1}\right)}{\det(\mathbf{A}^{-T}\mathbf{A}^{-1})} \right]. \end{aligned}$$

Ainsi, en omettant les termes qui ne dépendent pas de $\mathbf{A}$ dans le critère ci-dessus, on obtient le critère

$$C_2'(\mathbf{A}) = \sum_{i=1}^{M} \pi_m I(Y_1^{(m)}; \ldots; Y_N^{(m)}) + \frac{1}{2} \log_2 \left[\frac{\det \operatorname{diag}\left(\mathbf{A}^{-T}\mathbf{A}^{-1}\right)}{\det(\mathbf{A}^{-T}\mathbf{A}^{-1})} \right] \tag{6.6}$$

qui diffère du critère simplifié(5.11) du chapitre précédent par le fait que l'information mutuelle entre composantes $I(Y_1; \ldots; Y_N)$ a été remplacée par une moyenne pondérée des informations mutuelles entre composantes calculées pour chaque sous-bande.

On voit alors que même sous les hypothèses de données gaussiennes, la TKL, qui minimise le critère (5.11), n'est plus optimale en général pour le schéma séparable, car elle ne minimise pas le critère (6.6). Cette remarque est à l'origine de l'étude décrite dans [4].

6.5 Minimisation du critère dans le cas séparable

Les algorithmes `GCGsup` et `OrthICA` ont été modifiés pour donner les algorithmes `OST` (*Optimal Spectral Transform*) et `OrthOST` (*Orthogonal Optimal Spectral Transform*) qui retournent respectivement une matrice minimisant le critère du schéma séparable sans contrainte et avec la contrainte d'être orthogonale (pour plus de détails voir [5]). Le défaut majeur de ces algorithmes est de nécessiter un très grand nombre d'opérations arithmétiques (ils ont une complexité en $O(p(NrL + N^2L))$ où $r \in [30\,,\,60]$ et $p \in [20\,,\,60]$), rendant impossible leur utilisation dans un système embarqué à bord d'un satellite.

Nous verrons à la section 6.7 quelles approches nous avons explorées pour remédier à ce défaut majeur.

6.6 Résultats expérimentaux

Nous avons évalué les performances des transformations optimales sur deux familles d'images : trois images hyperspectrales, *Cuprite*, *Jasper* et *Moffett*, du capteur AVIRIS (gratuitement téléchargeable depuis le site http ://aviris.jpl.nasa.gov) et six images multispectrales, *Moissac*, *Port-de-Bouc*, *Vannes*, *Strasbourg*, *Montpellier*, *Perpignan*, fournies par le CNES et simulant des images que devrait fournir le nouveau capteur PLEIADES. Les images AVIRIS ont $N = 224$ composantes,

chacune de dimension 512×512 et sont codées sur 16 bits par pixel et par bande (bpppb). Les images PLEIADES ont toutes $N = 4$ composantes, elles sont toutes codées sur 12 bpppb et elles ont des dimensions variées (entre 200 et 300 colonnes pour quelques milliers de lignes). Les performances ont été évaluées en termes de distorsion finale versus débit total. Pour les images hyperspectrales, nous avons retenu quatre distorsions :

- l'erreur quadratique moyenne (norme 2) D exprimée en terme de rapport signal sur bruit (*Signal to Noise Ratio*) $\mathrm{SNR} = 10 \log_{10} \frac{\sigma^2}{D}$ où $\sigma^2 = \frac{1}{NL} \sum_{i=1}^{N} \sum_{n=1}^{L} (X_i(n) - \mu)^2$ (avec $\mu = \frac{1}{NL} \sum_{i=1}^{N} \sum_{n=1}^{L} X_i(n)$) est la variance empirique de l'image originale ;
- l'erreur absolue moyenne (norme 1 ou *Mean Absolute Error*) $\mathrm{MAE} = \frac{1}{NL} \sum_{i=1}^{N} \sum_{n=1}^{L} |X_i(n) - \widehat{X}_i(n)|$;
- l'erreur absolue maximale (norme infinie ou *Miximal Absolute Deviation*) $\mathrm{MAD} = \max\{|X_i(n) - \widehat{X}_i(n)| : 1 \leq i \leq N \text{ and } 1 \leq n \leq L\}$;
- l'erreur d'angle spectral maximale (*Maximum Spectral Angle*)

$$\mathrm{MSA} = \max \left\{ \operatorname{acos} \left(\frac{\sum_{i=1}^{N} X_i(n) \widehat{X}_i(n)}{\sqrt{\sum_{i=1}^{N} X_i^2(n) \sum_{i=1}^{N} \widehat{X}_i^2(n)}} \right) : 1 \leq n \leq L \right\} \tag{6.7}$$

  qui mesure l'écart maximal entre les spectres de l'image originale et de l'image reconstruite localisés à la même position spatiale et indépendamment de l'intensité lumineuse.

Ces quatre distorsions sont utiles et complémentaires pour estimer les performances d'un codeur d'images hyperspectrales dans diverses applications, comme la détection de cible ou la classification, d'après l'étude de E. Christophe, D. Léger et C. Mailhes, "Quality criteria benchmark for hyperspectral imagery", *IEEE Trans. Geoscience and Remote Sensing*, **43** (9), 2103–2114, 2005.

Pour les images multispectrales, nous n'avons considéré que deux distorsions : l'erreur quadratique moyenne D, exprimée en termes de rapport signal sur bruit crête (*Peak Signal to Noise Ratio*) $\mathrm{PSNR} = 10 \log_{10} \frac{(2^{N_b}-1)^2}{D}$, où N_b est le débit exprimé en bits par pixel et par bande de l'image originale ; et l'erreur absolue maximale.

Pour comparer différentes transformations dans le schéma séparable, nous avons utilisé le logiciel VM9 (*Verification Model version 9.1*) développé par le consortium JPEG2000 pour tester la partie 2 du standard. Le codeur entropique utilisé est donc EBCOT et une allocation optimale entre les différentes composantes et les différentes sous-bandes (en toute rigueur entre chaque *codeblock* est assurée par le VM9. De plus, le VM9 construit un vrai flot binaire contenant toute l'information utile au décodeur, en particulier les valeurs (codées en flottant sur 32 bits) des éléments de la matrice $\mathbf{A}^{-1}$ et les moyennes de chaque composante. Pour le réglage des paramètres, comme le nombre de niveaux de décompositions des DO 2D, nous avons choisi les valeurs par défaut du VM9 (à savoir 5 niveaux de décomposition).

Nous avons obtenu les résultats décrits dans les trois tableaux 6.1, 6.2 et 6.3. Nous avons pu constater que les transformations `OST` et `OrthOST` se comportaient comme on s'y attendait, conformément à la théorie. Nous avons également observé que, comme pour les images fixes 2D (voir chapitre 5), les performances des nouvelles transformations restaient très bonnes à moyens et bas débits et, pour les images hyperspectrales, que c'était vrai pour toutes les distorsions (ce n'est pas le cas pour les images multispectrales où l'erreur maximale absolue est plus grande avec `OrthOST`). Nous pouvons remarquer également que les performances de `OST` ne sont que très légèrement meilleures que celles d'`OrthOST`, ce qui peut s'xpliquer par le fait que la "distance à l'orthogonalité' est un terme pénalisant dans les critères d'optimalité. Le gain de codage généralisé des nouvelles transformations par rapport à la TKL n'est pas très important, mais quand même significatif : de l'ordre de 0,25 dB

	PSNR (dB)								MAD							
bit-rate	0,25	0,50	0,75	1,00	1,50	2,00	2,50	3,00	0,25	0,50	0,75	1,00	1,50	2,00	2,50	3,00
	Moissac															
Id	36,37	39,59	41,93	43,89	47,22	50,16	52,93	55,66	**691**	**366**	253	187	108	68	49	38
TKL	38,61	42,39	45,24	47,63	51,51	54,49	56,98	59,44	716	381	**214**	**135**	79	48	**32**	25
OrthOST	38,67	42,50	45,35	47,72	51,55	54,55	57,11	59,60	818	399	229	145	**78**	**47**	33	24
OST	**38,69**	**42,55**	**45,43**	**47,80**	**51,62**	**54,59**	**57,15**	**59,65**	745	496	215	138	**78**	48	**32**	**23**
	Port-de-Bouc															
Id	30,36	33,68	36,14	38,25	41,93	45,27	48,43	51,52	1198	653	544	361	198	135	85	64
TKL	**33,47**	37,74	40,88	43,45	47,53	50,89	53,82	56,53	922	**513**	**297**	**230**	139	**74**	**50**	35
OrthOST	33,42	37,80	41,05	43,71	47,90	51,31	54,28	56,99	885	**513**	305	237	**122**	77	51	**31**
OST	33,46	**37,85**	**41,12**	**43,78**	**48,00**	**51,40**	**54,36**	**57,06**	**866**	557	351	256	129	82	53	35
	Vannes															
Id	39,25	42,89	45,67	47,99	51,77	54,80	57,51	60,11	603	269	178	109	63	42	29	21
TKL	41,36	45,71	48,78	51,11	54,38	56,82	59,24	61,79	482	219	148	**86**	51	33	**24**	18
OrthOST	41,90	46,27	49,29	51,54	54,71	57,18	59,62	62,16	**354**	**190**	**135**	91	46	**30**	25	18
OST	**41,94**	**46,34**	**49,35**	**51,59**	**54,74**	**57,22**	**59,68**	**62,20**	393	204	138	88	**45**	33	25	**16**
	Strasbourg															
Id	30,82	34,19	36,73	38,91	42,70	46,09	49,20	52,13	1357	**877**	546	353	205	118	86	60
TKL	**32,51**	**36,59**	39,77	42,49	46,99	50,58	53,51	56,08	1041	927	**438**	403	184	90	52	**38**
OrthOST	**32,51**	**36,59**	39,78	42,50	47,01	50,61	53,55	56,11	**1010**	948	449	404	178	87	**50**	**38**
OST	32,49	**36,59**	**39,79**	**42,53**	**47,07**	**50,67**	**53,60**	**56,17**	1149	904	455	**289**	**162**	**81**	55	42
	Montpellier															
Id	32,17	35,23	37,59	39,62	43,17	46,30	49,17	51,95	1216	630	406	292	168	117	77	54
TKL	34,09	37,75	40,60	43,03	47,20	50,69	53,63	56,18	747	488	340	248	143	75	49	34
OrthOST	34,08	37,90	40,92	43,46	47,72	51,16	54,01	56,55	**681**	**454**	338	255	**127**	**68**	47	**32**
OST	**34,14**	**37,99**	**41,01**	**43,56**	**47,79**	**51,21**	**54,06**	**56,60**	704	483	**332**	**239**	**127**	**68**	**46**	33
	Perpignan															
Id	33,71	36,90	39,34	41,43	45,04	48,17	51,04	53,78	984	526	332	230	158	89	62	42
TKL	36,51	40,44	43,29	45,60	49,33	52,36	54,99	57,52	726	435	245	172	**84**	**54**	41	29
OrthOST	36,59	40,59	43,48	45,83	49,61	52,66	55,30	57,82	721	**371**	**232**	165	94	**54**	**37**	30
OST	**36,60**	**40,60**	**43,49**	**45,84**	**49,62**	**52,67**	**55,32**	**57,85**	**645**	383	292	**164**	94	55	38	**28**

TAB. 6.1 – Débit (en bpppb) versus PSNR (en dB) et versus MAD de différentes transformations spectrales sur des images multispectrales pour le schéma séparable (les meilleurs résultats sont en caractère gras). Le débit a été calculé avec le VM9.

pour les images multispectrales et de 0,45 dB pour les images hyperspectrales.

	SNR (dB)								MAE							
bit-rate	0,25	0,50	0,75	1,00	1,50	2,00	2,50	3,00	0,25	0,50	0,75	1,00	1,50	2,00	2,50	3,00
	Moffett															
Id	25,45	30,37	33,97	36,94	41,78	45,76	49,15	52,01	39,57	24,32	16,93	12,52	7,62	5,02	3,48	2,51
TKL	44,21	47,68	50,08	51,97	54,76	57,10	59,21	61,04	5,36	3,83	3,03	2,49	1,82	1,39	1,07	0,85
OrthOST	45,31	**48,57**	**50,87**	52,61	55,28	57,57	59,62	61,37	4,77	**3,47**	**2,78**	**2,32**	**1,72**	1,31	1,02	0,81
OST	**45,32**	48,56	**50,87**	**52,62**	**55,30**	**57,64**	**59,77**	**61,63**	**4,75**	**3,47**	**2,78**	**2,32**	**1,72**	**1,30**	**1,00**	**0,79**
	Cuprite															
Id	29,99	33,48	36,12	38,41	42,44	45,99	49,19	52,11	37,14	26,07	19,85	15,60	10,13	6,89	4,83	3,47
TKL	47,79	50,46	52,55	54,16	56,76	59,07	61,26	63,27	5,11	3,96	3,23	2,73	2,04	1,55	1,19	0,92
OrthOST	48,25	50,88	**52,89**	**54,44**	56,99	59,29	61,46	63,44	**4,83**	**3,79**	**3,12**	**2,65**	**1,98**	1,51	1,16	0,90
OST	**48,26**	**50,89**	**52,89**	**54,44**	**57,01**	**59,34**	**61,56**	**63,60**	**4,83**	**3,79**	**3,12**	**2,65**	**1,98**	**1,50**	**1,14**	**0,88**
	Jasper															
Id	21,34	24,83	27,56	29,92	34,01	37,67	41,09	44,33	64,84	45,61	34,39	26,82	17,23	11,52	7,89	5,49
TKL	42,93	46,49	48,61	50,37	53,18	55,56	57,72	59,66	5,78	4,04	3,27	2,72	1,99	1,51	1,16	0,91
OrthOST	43,66	46,94	49,02	50,73	53,47	55,81	57,94	59,85	5,35	3,85	3,13	2,62	1,93	1,46	1,13	0,88
OST	**43,70**	**46,96**	**49,05**	**50,74**	**53,50**	**55,87**	**58,03**	**60,01**	**5,32**	**3,84**	**3,12**	**2,61**	**1,92**	**1,45**	**1,11**	**0,87**

TAB. 6.2 – Débit (en bpppb) versus SNR (en dB) et versus MAE de différentes transformations spectrales sur des images hyperspectrales pour le schéma séparable. Le débit a été calculé ave le VM9.

Pour comparer entre eux les schémas de compression, puisque seul le séparable peut être testé par le VM9, nous avons estimé le débit au moyen de l'entropie d'ordre 1 moyenne :

$$\frac{1}{NP}\sum_{i=1}^{N}\sum_{\ell=1}^{P}\sum_{m=1}^{M}\pi_m H(\widehat{Y}_{i,\ell}^{(m)}),$$

où $H(\widehat{Y}_{i,\ell}^{(m)})$ est l'entropie d'ordre 1 de la variable quantifiée $\widehat{Y}_{i,\ell}^{(m)})$. Dans ce cas, la place mémoire occupée par la où les matrices spectrales inverses et les moyennes n'a pas été prise en compte.

Nous avons obtenu les résultats présentés dans le tableau 6.4.

	MSA (°)								MAD							
bit-rate	0,25	0,50	0,75	1,00	1,50	2,00	2,50	3,00	0,25	0,50	0,75	1,00	1,50	2,00	2,50	3,00
	Moffett															
Id	12,12	6,82	3,94	2,66	1,29	0,85	0,52	0,36	1676	781	492	1259	183	62	32	20
TKL	1,43	0,87	0,57	0,37	0,20	0,15	0,12	0,10	392	211	119	67	24	14	8	7
OrthOST	0,96	**0,47**	**0,31**	**0,25**	**0,18**	**0,14**	**0,11**	**0,09**	261	**77**	49	**33**	**18**	**10**	8	**6**
OST	**0,86**	0,50	0,32	**0,25**	**0,18**	**0,14**	**0,11**	**0,09**	**207**	101	**46**	37	19	12	**7**	**6**
	Cuprite															
Id	5,30	2,81	2,20	1,57	1,01	0,59	0,40	0,26	659	360	253	185	110	62	61	40
TKL	0,42	0,25	0,22	0,15	0,12	**0,08**	**0,07**	0,06	154	135	100	54	26	16	10	8
OrthOST	**0,32**	0,25	0,17	**0,14**	**0,10**	**0,08**	**0,07**	**0,05**	**113**	110	61	**37**	22	**11**	**9**	**7**
OST	0,35	**0,24**	**0,16**	**0,14**	**0,10**	**0,08**	**0,07**	**0,05**	**113**	**109**	**58**	42	**17**	**11**	**9**	**7**
	Jasper															
Id	18,20	12,53	7,88	5,70	3,87	2,14	1,41	1,01	1907	1220	732	559	241	160	84	55
TKL	0,91	0,53	0,43	0,34	0,26	0,20	0,15	0,12	225	151	82	57	30	15	10	7
OrthOST	0,83	**0,51**	**0,40**	0,33	**0,24**	0,19	0,15	0,12	157	**84**	**46**	**34**	23	**13**	9	7
OST	**0,79**	**0,51**	0,41	**0,32**	**0,24**	**0,18**	**0,14**	**0,11**	**156**	86	48	**34**	**22**	14	**8**	**6**

TAB. 6.3 – Débit (en bpppb) versus MSA (en degré °) et versus MAD de différentes transformations spectrales sur des images hyperspectrales pour le schéma séparable. Le débit a été calculé avec le VM9.

	SNR (dB)								MAD							
bit-rate	0,25	0,50	0,75	1,00	1,50	2,00	2,50	3,00	0,25	0,50	0,75	1,00	1,50	2,00	2,50	3,00
	Moffett															
TKL_1	44,47	47,95	50,43	52,37	55,26	57,90	60,48	63,07	325	206	102	65	29	16	10	6
TKL_2	45,35	48,57	50,99	52,82	55,66	58,28	60,85	63,42	340	191	113	71	33	17	9	**5**
TKL_3	45,47	48,75	51,21	53,06	55,99	58,69	61,29	63,88	346	250	190	129	88	53	19	8
$OrthOST_1$	45,53	48,77	51,16	52,95	55,75	58,35	60,91	63,48	**147**	81	61	47	**25**	**11**	**7**	6
$OrthICA_2$	45,74	48,92	51,30	53,06	55,86	58,47	61,03	63,60	169	**79**	**54**	49	26	16	9	**5**
$OrthICA_3$	45,83	49,04	51,44	53,25	56,17	58,89	61,55	64,18	305	185	104	59	34	17	10	6
OST_1	45,53	48,75	51,15	52,93	55,74	58,34	60,91	63,48	146	86	67	51	28	15	**7**	**5**
$GCGsup_2$	45,74	48,92	51,30	53,06	55,87	58,48	61,04	63,60	166	83	58	**45**	27	16	**7**	6
$GCGsup_3$	**45,88**	**49,09**	**51,47**	**53,28**	**56,21**	**58,94**	**61,62**	**64,27**	287	160	101	58	33	18	8	**5**
	Port de Bouc															
TKL_1	32,93	37,00	40,04	42,54	46,59	49,93	52,88	55,61	1021	549	362	246	147	91	70	35
TKL_2	33,01	37,09	40,16	42,65	46,73	50,08	53,07	55,85	924	557	379	246	157	91	54	36
TKL_3	33,25	37,28	40,30	42,81	46,88	50,24	53,23	56,01	947	685	388	304	171	97	53	38
$OrthOST_1$	32,91	37,03	40,20	42,76	46,93	50,32	53,30	56,06	935	625	351	249	145	85	68	36
$OrthICA_2$	33,87	37,49	40,40	42,86	46,97	50,44	53,41	56,10	**679**	**463**	**308**	272	140	**77**	50	35
$OrthICA_3$	34,09	37,82	40,67	43,08	47,18	50,62	53,58	56,27	763	524	386	269	**133**	91	49	35
OST_1	33,87	37,52	40,38	42,86	47,05	50,52	53,45	56,08	746	501	312	272	145	81	52	**34**
$GCGsup_2$	33,89	37,54	40,41	42,89	47,09	50,56	53,50	56,19	758	516	331	**245**	156	86	49	36
$GCGsup_3$	**34,10**	**37,86**	**40,71**	**43,17**	**47,33**	**50,76**	**53,68**	**56,36**	859	506	373	275	160	86	**47**	**34**

TAB. 6.4 – Débit (en bpppb) versus SNR pour *Moffett* ou PSNR pour *Port de Bouc* (en dB) et versus MAD de différentes transformations spectrales pour le schéma séparable ($_1$), en sous-bandes ($_2$) et mixte en sous-bandes ($_3$). Le débit a été estimé par l'entropie d'ordre 1 moyenne. Dans le schéma mixte en sous-bande, $P = 4$.

Pour les images hyperspectrales, nous avons observé un gain moyen d'environ 0,1 dB (resp. 0,85 dB) du schéma en sous-bandes par rapport au séparable pour les transformations `GCGsup` et `OrthICA` (resp. TKL) et un gain moyen d'environ 0,35 dB (resp. −0,dB) du schéma mixte en sous-bandes par rapport à celui en sous-bandes. Quand on tient compte de la taille du flot de bits occupé par les matrices spectrales inverses, il n'y a pas d'avantage à utiliser une variante non séparable, quelle qu'elle soit, du schéma séparable (sauf pour la TKL). Pour les images hyperspectrales, nous avons observé un gain moyen d'environ 0,06 dB (resp. 0,08 dB) du schéma en sous-bandes par rapport au séparable pour les deux transformations `GCGsup` et `OrthICA` (resp. la TKL) et un gain moyen d'environ 0,21 dB (resp. 0,15 dB) du schéma mixte en sous-bandes par rapport à celui en sous-bandes pour les transformations à base d'ACI (resp. la TKL). Quand on tient compte de la taille du flot de bits occupé par les matrices spectrales inverses, le léger avantage du schéma en sous-bandes sur celui séparable reste inchangé et celui du schéma mixte en sous-bandes sur celui en sous-bandes est réduit d'environ 0,1 dB.

6.7 Des transformations spectrales de moindre complexité

Nous avons vu que les algorithmes `OrthOST` et `OST` qui calculent les transformations spectrales optimales ont une grande complexité en nombre d'opérations. Pour remédier à ce défaut majeur, nous avons étudié les performances de transformations optimales exogènes, c'est-à-dire calculées grâce aux algorithmes ci-dessus sur une famille d'images issues d'un seul capteur (cette famille d'images constitue une base d'apprentissage) et appliquées sur d'autres images issues du même capteur. Cette étude a été financée par l'ESA (*European Space Agency*) dans le cadre d'une ITI et a été menée en partenariat avec la société LUXspace, qui a embauché Isidore Paul. Nous avons obtenu de très bonnes performances des transformations exogènes sur des images hyperspectrales Hyperion, cela a fait l'objet de trois présentations à des conférences avec actes et comité de lecture ([16, 14, 13]. De plus, compte-tenu des bonnes performances obtenues avec les transformations spectrales optimales exogènes, une demande de co-financement par l'ESA de la phase 3 (i.e., de l'implantation d'un prototype de codeur embarqué) de l'ITI HyperComp a été envoyée par LUXspace.

Sur une idée de Dinh Tuan Pham, nous avons exploré une autre voie pour réduire la complexité de la transformation spectrale optimale. L'idée était de modifier le critère à minimiser pour le schéma séparable en supposant les données gaussiennes. Le critère tenant compte des différences de variances des coefficients d'ondelettes par sous-bande, la transformation orthogonale qui le minimise n'est pas une TKL, tout en ne nécessitant que des estimations de moments d'ordres 1 et 2 (les estimations de statistiques, comme l'entropie différentielle, sont les étapes qui coûtent le plus d'opérations dans l'algorithme `OrthOST`). La minimisation du critère dans le cas gaussien peut se faire grâce à un algorithme de diagonalisation jointe de matrices (`JADO` développé par D. T. Pham. Ces recherches ont donné lieu à la publication d'un papier court dans la revue *Signal Processing* [4].

Enfin, nous avons également travaillé sur un codeur entropique à base d'arbres de zéros bien adapté aux transformations spectrales optimales et de faible complexité. Ceci, grâce au financement d'un post-doc de 2 ans par le CNES. Cette recherche a donné lieu aux publications suivantes : [15, 36, 43]. Les résultats obtenus sur des codeurs entropiques à base d'arbres de zéros bien adaptés aux transformations nouvelles à base d'ACI, couplés avec ceux menés sur les transformations exogènes ont fait l'objet de la rédaction d'un article soumis pour publication dans une revue scientifique [50].

6.8 Des transformations spectrales d'entiers en entiers

Signalons aussi que, depuis la rédaction de ce mémoire et ma demande d'inscription en HDR, nous avons commencé l'étude du comportement des transformations spectrales orthogonales exogènes en compression sans perte des images hyperspectrales. Pour cela, nous avons appliqué le procédé de Hao et Shi[1,2] qui consiste à factoriser une matrice orthogonale $\mathbf{A}$ sous la forme $\mathbf{A} = \mathbf{P}.\mathbf{L}.\mathbf{U}.\mathbf{S}$ où $\mathbf{P}$ est une matrice de permutation, $\mathbf{L}$ (resp. $\mathbf{U}$) une matrice triangulaire inférieure (resp. supérieure) avec des ± 1 sur la diagonale principale et $\mathbf{S}$ une matrice dont tous les éléments sont nuls, sauf sur la diagonale principale (où ils valent tous ± 1) et sur une ligne (où ils sont quelconques, en dehors de la diagonale). Grâce à cette factorisation, il est possible d'approcher la transformation $\mathbf{A}$ appliquée à un vecteur dont les composantes prennent des valeurs entières par une transformation réversible, dite d'entiers-en-entiers, (i.e., une injection de $\mathbb{Z}^n$ dans $\mathbb{Z}^n$ quand la matrice $\mathbf{A}$ est d'ordre n) appliquée au même vecteur. Cela permet un codage progressif en qualité pouvant atteindre le sans perte. La factorisation $\mathbf{PLUS}$ d'une matrice n'est pas unique et Hao et Shi proposent un algorithme pour choisir une telle factorisation qui, d'après les auteurs, est robuste face aux erreurs d'arrondi. Les premiers résultats obtenus en compression sans pertes d'images hyperspectrales par ce procédé appliqué à la transformation de Karhunen-Loève (exogène ou non) et aux transformations exogènes orthogonales que nous avons introduites ont été soumis pour publication dans un journal [51]. Nous pensons qu'ils peuvent être améliorés, car nous avons constaté sur des simulations que le conditionnement (pour la norme 2) des matrices $\mathbf{L}$, $\mathbf{U}$ et $\mathbf{S}$ peut être très grand, alors que celui de la matrice orthogonale $\mathbf{A}$ vaut 1. Dans de telles conditions les performances en compression de la transformation réversible sont médiocres (ce qui est normal, puisque la transformation réversible est alors éloignée de la transformation orthogonale optimale d'origine). Ainsi il pourrait être intéressant de laisser quelques degrés de liberté dans la factorisation $\mathbf{PLUS}$ de la matrice $\mathbf{A}$ (comme cela a été fait dans [26] pour des matrices de petites dimensions) pour rechercher une factorisation minimisant le conditionnement du produit $\mathbf{LUS}$. Une telle recherche pourra être très coûteuse en nombre d'opérations arithmétiques, sans augmenter la complexité de codecs basés sur de telles transformations réversibles exogènes, car elle sera faite une fois pour toutes sur une base d'apprentissage. C'est une perspective possible, permettant de mieux comprendre pourquoi les premiers résultats obtenus en compression sans perte avec les transformations orthogonales exogènes réversibles que nous avons introduites ne sont pas tellement meilleurs qu'avec les transformations de Karhunen-Loève exogènes réversibles.

[1] [Hao-Shi–2000] P. Hao et Q. Shi, "Matrix factorization for reversible integer mapping", *IEEE Transactions on Signal Processing*, vol. 49, no. 10, pp. 2314–2324, 2000.

[2] [Hao-Shi–2003] P. Hao et Q. Shi, "Reversible integer KLT for progressive-to-lossless compression of multiple component images", *Proceedings of the IEEE International Conference on Image Processing ICIP'03*, 2003.

Chapitre 7

Conclusion et perspectives

Nous avons vu dans ce mémoire, avec les thèses de Michel Narozny et Isidore Paul Akam Bita, comment nous avons pu améliorer les codeurs d'images multi-composantes existants. Pour cela, nous avons commencé par mieux comprendre les liens entre compression de données et analyse en composantes indépendantes, grâce à une modélisation mathématique adéquate, où l'hypothèse de données gaussiennes est relâchée. Ce nouveau point de vue a permis de modifier un algorithme d'analyse en composantes indépendantes existant pour qu'il réponde efficacement au problème de la compression de données : sous certaines hypothèses réalistes peu restrictives, la transformation retournée est optimale à hauts débits en codage par transformée. Nous avons ensuite recherché des images bien adaptées à ces nouvelles transformations, passant des images en niveaux de gris, à celles en couleur et aux images multi- et hyper-spectrales. Nous avons dû repasser par l'étape de modélisation quand nous avons associé transformation linéaire pour réduire la redondance spectrale et décomposition en ondelettes 2D pour réduire la redondance spatiale. Puis nous avons modifié les transformations et les algorithmes qui les calculent, pour mieux coller aux images et ... finalement trouver des outils plus efficaces que l'état de l'art en codage d'images hyperspectrales.

Enfin, grâce à des techniques d'apprentissage, nous avons proposé des transformations exogènes légèrement sous-optimales mais encore très performantes. Cela a permis de réduire considérablement la complexité des codecs basés sur les nouvelles transformées et, à nos partenaires, d'envisager de les embarquer à bord de satellites. Cette dernière étape de réduction de complexité est le fruit de travaux réalisés après les thèses mentionnées ci-dessus, grâce aux financements de l'ESA (ITI HyperComp) et du CNES (post-doc de Mohamed Hariti). Les résultats obtenus dans cette dernière étape ont donné lieu à la rédaction de deux articles [50, 51] soumis pour publication dans des revues internationales. Il reste encore à étudier le problème de la généralisation, c'est-à-dire celui de quantifier et de borner les erreurs ou les baisses de performances quand l'outil "appris" (par exemple la transformation spectrale) est appliqué à de nouvelles données. Ce problème est voisin de celui dit d'extrapolation en résolution numérique d'EDP et intéresse la communauté mathématique, en particulier Jean-Pierre Croisille, Professeur de mathématiques à l'Université de Metz. J'aimerais en savoir plus sur la théorie de la généralisation (ou extrapolation) de l'apprentissage statistique. Cela pourra éventuellement s'appliquer aux transformations exogènes mentionnées ci-dessus.

Avant de présenter mes perspectives de recherche, je dois préciser mon environnement de travail, qui est en pleine évolution. Depuis deux ans je suis rattaché à l'équipe de recherche Information, Multimodalité & Signal (IMS) de Supélec, basée sur son campus de Metz. C'est une équipe propre à l'école qui a l'ambition à court terme d'être équipe d'accueil reconnue par le Ministère. La compression d'images ou de données n'est pas un axe de recherche de l'équipe, et je dois donc progressivement changer de thématique. Voici le descriptif des problèmatiques de l'équipe IMS (extrait de son site

internet) :

> « Dans de nombreux domaines applicatifs, les phénomènes faisant l'objet d'une analyse ou d'un traitement se manifestent par le biais de différents supports d'information. Ces différents supports sont parfois porteurs d'informations complémentaires et indépendantes. Dans ce cas, la prise en compte simultanée des différentes observations peut être nécessaire à la compréhension totale du processus générateur. Dans d'autres cas les différents supports portent des informations redondantes à des degrés divers et leur étude globale permet, outre une compréhension complète, d'assurer une plus grande robustesse de l'analyse. L'extraction d'une information pertinente à partir de flux dissociés d'informations brutes complémentaires et/ou redondantes de natures hétérogènes peut-être vue comme l'utilisation de plusieurs modalités dans un processus d'analyse, ce qui justifie l'appellation d'analyse multimodale ou de traitement multimodal de l'information.
>
> Grâce à l'évolution de la puissance de calcul, de la capacité de stockage et des bandes passantes des réseaux informatiques on peut aujourd'hui appréhender le traitement de l'information sous l'angle de la multimodalité et de l'analyse de données. L'aspect multimodal est alors à prendre au sens large, faisant intervenir des informations provenant de sources différentes mais aussi de natures hétérogènes. Le traitement multimodal de l'information peut dès lors nécessiter l'accès à des masses conséquentes de données de natures différentes - numériques et/ou symboliques - disponibles sous formes de bases statiques (bases de données) ou de flux (i.e. de type perceptuel) et qu'il faut considérer comme porteuses d'informations sur les signaux et systèmes qui en sont à l'origine. Le but de la recherche présentée ici consiste alors à extraire les informations pertinentes de ces données hétérogènes pour créer des modèles capables d'analyser (modélisation, identification), reproduire (simulation), classifier (détection, reconnaissance de formes) ou contrôler (automatique, robotique, interfaces homme-machine) le comportement des systèmes qui ont généré les données ou même d'en améliorer la qualité (filtrage, restauration) ou de les compresser. L'information quantitative (valeurs numériques, paramètres de modèles, etc.) et l'information symbolique (existences, représentation sémantique, structures de modèles, etc.) doivent pouvoir être couplées pour aboutir à de meilleures performances. Ce couplage numérique/symbolique fera partie des axes forts de recherche. Les groupes de compétences décrits ci-après contribuent à la recherche sur ce thème. »

Parmi les sujets et action de recherche de l'équipe, je peux plus facilement m'intégrer à l'axe "Traitement statistique du signal et apprentissage numérique" décrit sur le site de l'équipe en les termes suivants :

> « Le traitement du signal et plus particulièrement le traitement statistique du signal offre une grande variété de techniques. Ces techniques s'appliquent essentiellement aux informations de type analogique ou numérique comme le son, l'image et la vidéo. C'est sous l'angle du traitement de l'information qu'est abordé le problème. En particulier, l'étude et la mise au point de méthodes d'analyse en composantes indépendantes font l'objet de recherches au sein de l'équipe. Les méthodes d'apprentissage ou d'intelligence artificielle numérique et probabiliste (réseaux de neurones, SVM - Support Vector Machines -, Processus décisionnels de Markov) constituent d'autres voies. Du point de vue de la théorie de l'information toutes ces méthodes peuvent s'unifier et c'est sous cet angle que le traitement multimodal de l'information est abordé. »

Dans ce contexte, j'aimerais étudier les problèmes d'estimation pure et de détection, sans connaissance *a priori* de la vraisemblance (c'est-à-dire de la densité de probabilité de l'observation, qui

dépend du paramètre à estimer). Ce sont des hypothèses réalistes, rencontrées dans de nombreuses situations réelles. Il s'agit alors d'estimer la vraisemblance quand c'est possible, ou de faire quand même quelque chose sans cette connaissance *a priori*. Ce problème est étudié depuis quelques décennies et j'aimerais en savoir plus sur le sujet, situé à l'intersection des statistiques et de l'intelligence artificielle. Une première étape consistera à étudier l'estimation de densités de probabilités, en particulier pour des observations vectorielles. Par le biais d'un projet d'élèves de 3ème année, nous avons commencé l'étude de l'estimation des vraisemblances (ou plus exactement de quelques unes de leurs lois marginales) dans un espace de dimension 16 (espace des cepstres) pour, peut-être, proposer de nouvelles approches au problème de détection de changement de locuteurs. Cette étude est co-encadrée par Jean-Louis Gutzwiller et Stéphane Rossignol, membres de l'équipe IMS. Pour la connaissance que j'en ai aujourd'hui, il me semble que les méthodes proposées sont optimales sous certaines hypothèses, dont la gaussianité des données et se posent alors les questions "Y-a-t-il un intérêt pratique à relâcher l'hypothèse de gaussianité ? et "Est-ce qu'on peut le faire ?" Il s'agit bien sûr de poser et éventuellement de répondre à ces questions dans un cadre mathématique rigoureux.

Par ailleurs, comme cela a déjà été écrit dans l'introduction de ce mémoire, les résultats présentés dans la deuxième partie de la thèse d'Isidore Paul Akam Bita n'ont pas encore eu la publicité qu'ils méritent. Cette partie, qui consiste à appliquer un modèle convolutif pour de la compression de données en relâchant l'hypothèse de gaussianité, est la plus innovante du travail de thèse de Paul. L'aspect mathématique est déjà correctement rédigé dans [49], il montre des liens prometteurs avec la théorie des bancs de filtres optimaux en codage. En revanche, toute la partie programmation est à revoir (l'implantation réalisée par I.P. Akam Bita nécessite énormément de mémoire et ne permet pas, de ce fait, de mettre en valeur la méthode). J'aimerais poursuivre ces recherches en suivant la même démarche que celle rappelée ci-dessus, à savoir :

1. description de critères à minimiser pour construire des transformations optimales en codage d'images mono- et multi-composantes sous des hypothèses réalistes (données non gaussiennes, codage séparé des composantes, quantification à haute résolution, allocation optimale),
2. implantation d'algorithmes efficaces pour minimiser ces critères,
3. comparaison des performances entre différents schémas de compression,
4. puis apprentissage de transformations exogènes quasi-optimales.

Les étapes 1 et 2 ont déjà été abordées dans la thèse de Paul. La rédaction du papier en préparation [49] montre que notre approche est une extension de la théorie des bancs de filtres bi-orthogonaux optimaux en codage en ayant relâché l'hypothèse de données gaussiennes.

Les principales difficultés à surmonter seront dans un premier temps l'implantation d'algorithmes efficaces de séparation/déconvolution adaptés aux critères rencontrés en compression de données. Remarquons que contrairement au cas de mélange instantané, l'analyse en composantes indépendantes avec mélange convolutif ne dispose pas encore aujourd'hui (à ma connaissance) d'algorithmes efficaces et robustes ayant faits leurs preuves dans diverses applications. Ainsi, cette première phase sera plus difficile à surmonter que dans le cas d'un mélange instantané.

Je suis persuadé qu'il y a matière pour d'intéressantes recherches dans cette direction, sous réserve de trouver des financements. J'espère en obtenir en continuant de travailler avec le CNES à Toulouse et la société LuxSpace au Luxembourg. Signalons ici que la quasi totalité de la littérature sur les bancs de filtres à reconstruction parfaite optimaux en codage repose sur l'hypothèse de données gaussiennes (voir par exemple le rapport de thèse d'Hocine Bekkouche). Les premiers résultats obtenus avec Isidore Paul Akam Bita et Dinh-Tuan Pham [49] permettent de généraliser un théorème caractérisant des bancs de filtres bi-orthogonaux optimaux en relâchant l'hypothèse de gaussianité. S'ouvre ainsi

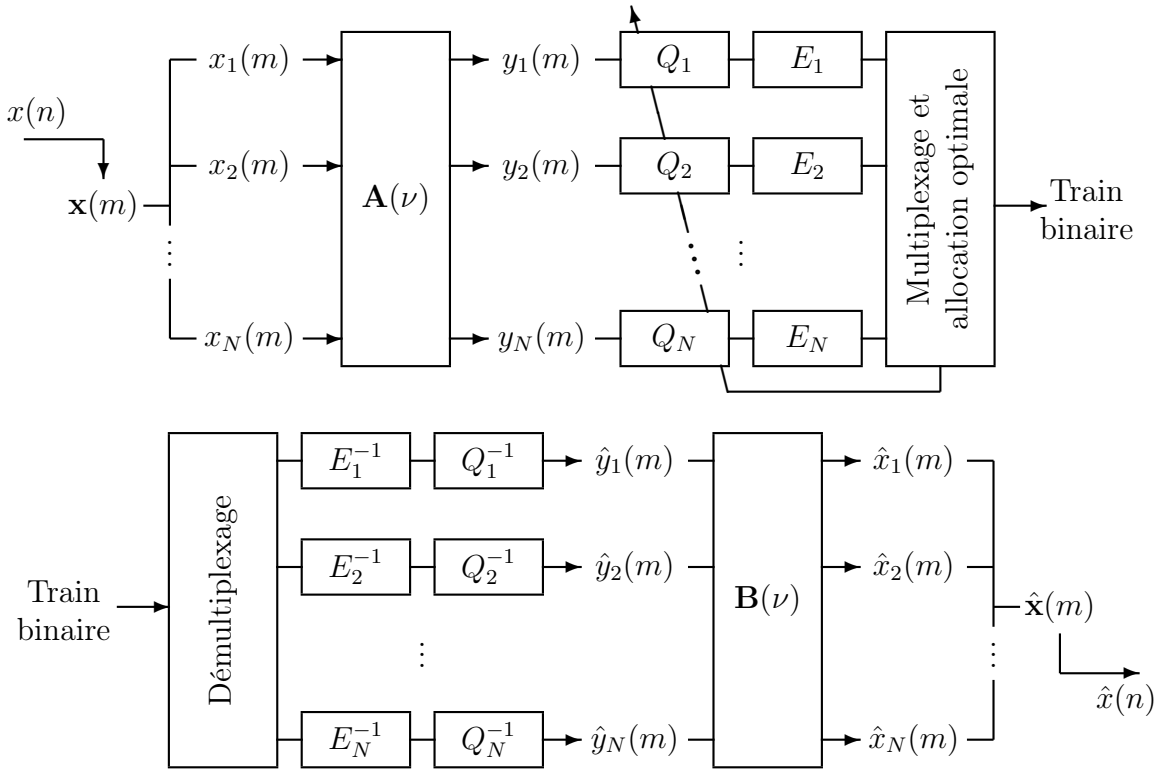


FIG. 7.1 – Codeur/décodeur par banc de filtres ($\mathbf{A}(\nu)$ et $\mathbf{B}(\nu)$ sont les réponses fréquentielles respectives du banc de filtres d'analyse et de celui de synthèse). Pendant le codage le quantificateur Q_i suivi du codeur entropique E_i est appliqué à la $i^{\text{ème}}$ composante transformée y_i. Pendant le décodage le décodeur entropique E_i^{-1} (inverse mathématique de E_i) suivi de la déquantification Q_i^{-1} (qui n'est pas l'opération mathématique inverse de Q_i) est appliqué pour reconstruire une approximation $\hat{y}_i$ de la $i^{\text{ème}}$ composante transformée.

un axe de recherche qui, si les calculs ne sont pas trop ardus, peut s'avérer très intéressant, aussi bien d'un point de vue pratique que d'un point de vue théorique.

Bien sûr, d'autres questions m'intéressent, par exemple, dans l'optique de mieux comprendre les systèmes naturels de codage d'information (comme dans les cerveaux) : quelle est l'influence de la fonction coût servant à définir la distorsion que l'on cherche à minimiser (pour une taille de flot de bits donnée) dans le cas de la compression avec pertes ? Nous avons vu que dans le cas d'un coût quadratique, un critère à minimiser se décompose en la somme de l'information mutuelle entre composantes transformées et d'un terme positif qui peut s'interpréter comme une sorte de distance à l'orthogonalité de la transformation. Le fait que les transformations orthogonales soient privilégiées est dû au choix du coût quadratique. Mais il n'y a pas de raison (à ma connaissance) pour que la Nature privilégie le coût quadratique par rapport à un autre. D'où la question, et si on change de coût, comment sont modifiés les critères ? Nous avons vu également comment les avancées réalisées ces dernières années en ACI ont pu enrichir les connaissances des systèmes de compression de données. La question de l'enrichissement inverse se pose. Et la réponse à la question précédente pourrait y participer ...

Pour conclure, il est important de remarquer que les recherches d'algorithmes efficaces d'ACI orthogonale et de séparation/déconvolution unitaire, c'est-à-dire préservant l'énergie, trouvent beaucoup d'applications autres que la compression de données. En effet, ce seront les mêmes que celles de l'analyse en composantes principales des cours de statistique ou de la "vraie" décomposition de Karhunen-Loève (enseignée dans les cours de statistiques, et pas la version faible des cours de compression de données). Indiquons à titre d'exemple la décomposition de solutions, en régime turbulent stationnaire, d'EDP (équation d'écoulement de Navier et Stockes) en sommes d'éléments plus simples à étudier (voir par exemple "Intermodal energy transfers in a proper orthogonal decomposition-Galerkin representation of a turbulent separated flow" de M. Couplet, P. Sagaut et C. Basdevant, *J. Fluid Mech.*, vol. 491, pp. 275–284, 2003). La résolution des EDP est un des sujets de recherche de l'équipe "mathématiques appliquées" du LMAM (Laboratoire de Mathématiques et Applications de Metz). Il est prévu que j'aille présenter mes travaux de recherche lors d'un séminaire de mathématiques. J'ai des contacts depuis de nombreuses années avec des professeurs du département de mathématiques de l'Université Paul Verlaine (UPV) de Metz, en particulier pour le master recherche (ex DEA) de mathématiques fondamentales et appliquées cohabilité entre l'UPV et Supélec. L'obtention d'une HDR devrait faciliter les collaborations au niveau doctorat que je pourrai avoir avec le LMAM. Il existe plusieurs axes potentiels. Par exemple, Ralph Chill, Professeur de mathématiques et membre du LMAM avec qui j'ai co-encadré plusieurs stages de Master, s'intéresse depuis plusieurs années à des applications en traitement d'images des EDP. Mais celui mentionné juste au-dessus : étude des transformées linéaires (modèles instantané et convolutifs) conservant l'énergie et adaptant au mieux des données particulières à leur codage, dans des situations réelles d'écoulements (équations de Navier-Stocke ou acoustique automobile[1]), est celui qui a ma préférence aujourd'hui et il pourrait intéresser Jean-Pierre Croisille du LMAM.

[1]Voir l'étude que j'ai réalisés avec Renault dans mon CV, page 11.

Annexe A

Éléments de la théorie de l'information et de la quantification

A.1 Courbes débit vs distorsion d'une source d'information sans mémoire

Nous rappelons dans cette annexe les conditions de validation et l'énoncé du théorème de codage de Shannon. Pour la $i^{\text{ème}}$ composante et une taille $V > 0$ donnée, considérons une distorsion

$$\begin{array}{rcl} d_i \,:\, \mathbb{R}^V \times \mathbb{R}^V & \to & \mathbb{R}_+ \\ (\mathbf{y}_i^V, \hat{\mathbf{y}}_i^V) & \mapsto & d_i(\mathbf{y}_i^V, \hat{\mathbf{y}}_i^V), \end{array}$$

une fonction de codage $f_{i,1}$ et une fonction de décodage $g_{i,1}$:

$$f_{i,1} \,:\, \mathbb{R}^V \to [\![1\,;\,2^{R_i}]\!] \qquad \text{et} \qquad g_{i,1} \,:\, [\![1\,;\,2^{R_i}]\!] \to \mathbb{R}^V, \tag{A.1}$$

posons $\hat{\mathbf{Y}}_i^V = g_{i,1}\left[f_{i,1}\left(\mathbf{Y}_i^V\right)\right]$. Plus généralement, pour une famille finie $\mathbf{y}_i^{nV} = (\mathbf{y}_{i,1}^V, \ldots, \mathbf{y}_{i,n}^V)$ de n réalisations de $\mathbf{Y}_i^V$, pour la fonction distorsion

$$d_i(\mathbf{y}_i^{nV}, \hat{\mathbf{y}}_i^{nV}) = \frac{1}{n}\sum_{k=1}^{n} d_i(\mathbf{y}_{i,k}^V, \hat{\mathbf{y}}_{i,k}^V) \qquad \text{où} \qquad \hat{\mathbf{y}}_i^{nV} = (\hat{\mathbf{y}}_{i,1}^V, \ldots, \hat{\mathbf{y}}_{i,n}^V), \tag{A.2}$$

et pour des fonctions de codage $f_{i,n}$ et de décodage $g_{i,n}$:

$$f_{i,n} \,:\, \left(\mathbb{R}^V\right)^n \to [\![1\,;\,2^{nR_i}]\!] \qquad \text{et} \qquad g_{i,n} \,:\, [\![1\,;\,2^{nR_i}]\!] \to \left(\mathbb{R}^V\right)^n \tag{A.3}$$

posons $\hat{\mathbf{Y}}_i^{nV} = g_{i,n}\left[f_{i,n}\left(\mathbf{Y}_i^{nV}\right)\right]$.

En suivant la présentation de[1] pages 340–342, on dit d'un *couple débit-distorsion* $(R_i, D_i) \in \mathbb{R}_+^2$ qu'il est *réalisable* quand il existe une suite $(f_{i,n}, g_{i,n})_{n\geq 1}$ de fonctions de codage et de décodage associées au même débit R_i pour laquelle

$$\lim_{n\to\infty} \mathrm{E}\left[d_i(\mathbf{Y}_i^{nV}, \hat{\mathbf{Y}}_i^{nV})\right] \leq D_i. \tag{A.4}$$

Pour toute distorsion $D_i > 0$ on pose par définition

$$R_i^{\mathrm{Sh}}(D_i) = \inf\{R_i \in \mathbb{R}_+ \,:\, (R_i, D_i) \text{ est un couple débit-distorsion réalisable}\} \tag{A.5}$$

[1] T. M. Cover et J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.

et le théorème de codage de Shannon (théorème[2] 13.2.1, page 342 du livre de Cover et Thomas) affirme que pour une suite indépendante et identiquement distribuée $(\mathbf{Y}_{i,k}^V)_{k\geq 1}$ on a

$$R_i^{\mathrm{Sh}}(D_i) = \min\left\{ I(\mathbf{Y}_i^V;\hat{\mathbf{Y}}_i^V) \,:\, \mathrm{E}\left[d_i(\mathbf{Y}_i^V,\hat{\mathbf{Y}}_i^V)\right] \leq D_i \right\}. \tag{A.6}$$

Le minimum de l'information mutuelle $I(\mathbf{Y}_i^V;\hat{\mathbf{Y}}_i^V)$ est obtenue pour une densité de probabilité conditionnelle $p(\hat{\mathbf{y}}_i^V \,|\, \mathbf{y}_i^V)$ particulière, correspondant à l'optimum.

Dans le cas d'une distorsion quadratique :

$$d_i(\mathbf{y}_i^V,\hat{\mathbf{y}}_i^V) = \|\mathbf{y}_i^V - \hat{\mathbf{y}}_i^V\|_2^2 = \sum_{k=1}^{V}(y_{i,k}-\hat{y}_{i,k})^2 \tag{A.7}$$

on montre[3] que pour une faible distorsion D_i

$$R_i^{\mathrm{Sh}}(D_i) \simeq h(\mathbf{Y}_i^V) - \frac{V}{2}\log_2\left[\frac{2\pi e D_i}{V}\right]. \tag{A.8}$$

En changeant d'unités, c'est-à-dire en exprimant le débit R_i en bits par échantillon et en prenant une distorsion moyenne par échantillon D_i/V la relation précédente devient

$$R_i^{\mathrm{Sh}}(D_i) \simeq \frac{h(\mathbf{Y}_i^V)}{V} - \frac{1}{2}\log_2[2\pi e D_i]. \tag{A.9}$$

A.2 Quantificateur vectoriel à haute résolution

Nous considérons un quantificateur vectoriel Qv_i appliqué à la $i^{\text{ème}}$ composante $\mathbf{Y}_i^V$. Nous supposons que le *quantificateur vectoriel* est *en réseau*, c'est-à-dire avec des cellules élémentaires qui

[2]Dans le livre de Cover et Thomas (indiqué à la note du bas de la page 69), les théorèmes sont présentés dans le cas de VA discrètes avec une distorsion bornée, mais il est indiqué (page 344, début du paragraphe 13.3.2) qu'elles s'étendent au cas de VA continues et de distorsion non bornée.

[3]En suivant le raisonnement donné dans le livre de Cover et Thomas page 345, on a

$$\begin{aligned}
I(\mathbf{Y}_i^V;\hat{\mathbf{Y}}_i^V) &= h(\mathbf{Y}_i^V) - h(\mathbf{Y}_i^V \,|\, \hat{\mathbf{Y}}_i^V) = h(\mathbf{Y}_i^V) - h(\mathbf{Y}_i^V - \hat{\mathbf{Y}}_i^V \,|\, \hat{\mathbf{Y}}_i^V) \\
&\geq h(\mathbf{Y}_i^V) - h(\mathbf{Y}_i^V - \hat{\mathbf{Y}}_i^V) \geq h(\mathbf{Y}_i^V) - h(\mathcal{N}(\underline{0},\mathbf{K}_{\mathbf{Y}_i^V-\hat{\mathbf{Y}}_i^V})) \\
&\qquad \text{où} \quad \mathbf{K}_{\mathbf{Y}_i^V-\hat{\mathbf{Y}}_i^V} = \mathrm{E}\left[\left(\mathbf{Y}_i^V-\hat{\mathbf{Y}}_i^V\right)\left(\mathbf{Y}_i^V-\hat{\mathbf{Y}}_i^V\right)^T\right] \\
&= h(\mathbf{Y}_i^V) - \frac{V}{2}\log_2[2\pi e] - \frac{1}{2}\log_2\left[\det \mathbf{K}_{\mathbf{Y}_i^V-\hat{\mathbf{Y}}_i^V}\right] \\
&\geq h(\mathbf{Y}_i^V) - \frac{V}{2}\log_2[2\pi e] - \frac{V}{2}\log_2\left[\mathrm{tr}\left(\mathbf{K}_{\mathbf{Y}_i^V-\hat{\mathbf{Y}}_i^V}\right)/V\right] \\
&= h(\mathbf{Y}_i^V) - \frac{V}{2}\log_2\left[\frac{2\pi e D_i}{V}\right],
\end{aligned}$$

où $\mathcal{N}(\underline{0},\mathbf{K})$ désigne un vecteur aléatoire gaussien de moyenne $\underline{0}$ et de matrice variance-covariance $\mathbf{K}$. En étudiant les inégalités ci-dessus et les conditions pour lesquelles elles deviennent des égalités, on vérifie qu'en introduisant un vecteur aléatoire $\mathbf{Z}_i^V$ de dimension V, gaussien, centré, de matrice variance-covariance $\frac{D_i}{V}\mathbf{I}_V$ (où $\mathbf{I}_V$ est la matrice identité d'ordre V) et indépendant de $\mathbf{Y}_i^V$, et en posant $\hat{\mathbf{Y}}_i^V = \mathbf{Y}_i^V + \mathbf{Z}_i^V$, on a

$$I(\mathbf{Y}_i^V;\hat{\mathbf{Y}}_i^V) \simeq h(\mathbf{Y}_i^V) - \frac{V}{2}\log_2\left[\frac{2\pi e D_i}{V}\right] \qquad \text{et} \qquad \mathrm{E}\left[\left\|\mathbf{Y}_i^V-\hat{\mathbf{Y}}_i^V\right\|_2^2\right] = D_i$$

pour une distorsion D_i faible (pour laquelle on écrit $h(\mathbf{Z}_i^V \,|\, \hat{\mathbf{Y}}_i^V) \simeq h(\mathbf{Z}_i^V \,|\, \mathbf{Y}_i^V) = h(\mathbf{Z}_i^V)$).

forment un pavage régulier de $\mathbb{R}^V$ (en fait comme nous le verrons ci-dessous, d'un sous-ensemble borné de $\mathbb{R}^V$) et d'avoir un nombre fini K_i de cellules élémentaires notées $\mathrm{cel}(k)$ $(1 \leq k \leq K_i)$. Notons $\hat{\mathbf{z}}_i^V(k)$ la valeur de sortie du quantificateur quand $\mathbf{Y}_i^V \in \mathrm{cel}(k)$:

$$\mathrm{Qv}_i[\mathbf{Y}_i^V] = \hat{\mathbf{z}}_i^V(k) \quad \Leftrightarrow \quad \mathbf{Y}_i^V \in \mathrm{cel}(k) \qquad (1 \leq k \leq K_i).$$

Posons

$$\hat{\mathbf{Y}}_i^V = \mathrm{Qv}_i[\mathbf{Y}_i^V] = (\hat{X}_{i,1}, \ldots, \hat{X}_{i,V}) \quad \text{et} \quad \hat{\mathbf{Y}}^V = \mathrm{Qv}[\mathbf{Y}^V] = \begin{pmatrix} \mathrm{Qv}_1[\mathbf{Y}_1^V] \\ \vdots \\ \mathrm{Qv}_N[\mathbf{Y}_N^V] \end{pmatrix}.$$

Le pavage régulier de $\mathbb{R}^V$ est obtenu en appliquant une transformation affine (i.e., une transformation linéaire $\mathbf{U}_i : \mathbb{R}^V \to \mathbb{R}^V$ et une translation) inversible au réseau cubique $\delta_i \mathbb{Z}^V$ formé d'hyper-cubes de $\mathbb{R}^V$ de longueur de côté $\delta_i > 0$.

Le vecteur aléatoire $\mathbf{Y}_i^V$ est supposé admettre une densité de probabilité (ddp) continue sur $\mathbb{R}^V$ et nous supposons que le *quantificateur vectoriel* Qv_i est *à haute résolution*, c'est-à-dire vérifie les trois conditions suivantes.

1. Il existe un pavé borné $\pi_i = [a_{i,1}\,;\, b_{i,1}] \times \cdots \times [a_{i,V}\,;\, b_{i,V}] \subset \mathbb{R}^V$ tel que $P(\mathbf{Y}_i^V \notin \pi_i) \simeq 0$.
2. Les cellules élémentaires sont suffisamment petites par rapport aux fluctuations de la ddp de $\mathbf{Y}_i^V$ pour que $\forall k \in [\![1\,;\, K_i]\!]$, $\exists p_{i,k} > 0$ tel que $\forall \mathbf{y}_i^V \in \mathrm{cel}(k)$, $p(\mathbf{y}_i^V) \simeq p_{i,k}$.
3. $\forall k \in [\![1\,;\, K_i]\!]$, la valeur de sortie de la $k^{\text{ème}}$ cellule coïncide avec le centre de gravité de la cellule : $\hat{\mathbf{z}}_i^V = \left(\int_{\mathrm{cel}(k)} \mathbf{y}^V \, d\mathbf{y}^V\right) / \mathrm{vol}(k)$, où $\mathrm{vol}(k) = \int_{\mathrm{cel}(k)} d\mathbf{y}^V$ est le volume de la cellule.

Le quantificateur étant en réseau, le volume de la $k^{\text{ème}}$ cellule et son moment d'inertie par rapport à son centre de gravité sont indépendants de k. Nous notons $\delta_i^{V+2} M(\mathbf{U}_i)$ le moment d'inertie d'une cellule élémentaire par rapport à son centre de gravité. La grandeur $M(\mathbf{U}_i)$ ne dépend que de la forme du réseau (de $\mathbf{U}_i$), elle ne dépend pas du pas de quantification δ_i, et le volume d'une cellule élémentaire vaut $\delta_i^V |\det \mathbf{U}_i|$.

A.3 Quantificateur scalaire à haute résolution

Dans ce cas, chaque composante de $\mathbf{Y}$ est quantifiée avec un quantificateur scalaire régulier. Celui de la $i^{\text{ème}}$ composante est noté Qs_i, avec $\eta_{i,1}$, $\eta_{i,2}$, …, η_{i,K_i-1} ses points frontières et $\hat{z}_{i,1}$, $\hat{z}_{i,2}$, …, $\hat{z}_{i,K_i}$ ses sorties :

$$\mathrm{Qs}_i[Y_i] = \begin{cases} \hat{z}_{i,1} & \Leftrightarrow & Y_i \leq \eta_{i,1} \\ \hat{z}_{i,k} & \Leftrightarrow & \eta_{i,k-1} < Y_i \leq \eta_{i,k} \quad (k \in [\![2\,;\, K_i - 1]\!]) \\ \hat{z}_{i,K_i} & \Leftrightarrow & \eta_{i,K_i-1} < Y_i. \end{cases} \tag{A.10}$$

Nous avons $\hat{Y}_i = \mathrm{Qs}_i[Y_i]$ et nous notons $\mathcal{Z}_i = \{\hat{z}_{i,1}, \hat{z}_{i,2}, \ldots, \hat{z}_{i,K_i}\}$ l'ensemble des valeurs de sortie du quantificateur. Le résultat de la quantification appliquée au vecteur $\mathbf{Y}$ est noté $\hat{\mathbf{Y}} = \mathrm{Qs}[\mathbf{Y}] = (\hat{Y}_1, \ldots, \hat{Y}_N)^T$. Nous nous intéressons au codage de séquences finies de vecteurs aléatoires discrets $\hat{\mathbf{Y}}^V = (\hat{\mathbf{Y}}_1, \hat{\mathbf{Y}}_2, \ldots, \hat{\mathbf{Y}}_V)$, où pour tout entier k, $\hat{\mathbf{Y}}_k$ est le résultat de la quantification appliquée au vecteur $\mathbf{Y}_k$: $\hat{\mathbf{Y}}_k = \mathrm{Qs}[\mathbf{Y}_k]$. La $i^{\text{ème}}$ composante du vecteur $\hat{\mathbf{Y}}_k$ est notée $\hat{Y}_{i,k}$ et la séquence finie des $\hat{Y}_{i,k}$ (resp. $Y_{i,k}$) obtenue en faisant varier k est notée $\hat{\mathbf{Y}}_i^V = (\hat{Y}_{i,1}, \hat{Y}_{i,2}, \ldots, \hat{Y}_{i,V})$ (resp. $\mathbf{Y}_i^V = (Y_{i,1}, \ldots, Y_{i,V})$).

Nous supposons que chaque composante de $\mathbf{Y}$ est quantifiée à *haute résolution*[4], c'est-à-dire que pour tout $i \in [\![1\,;\,N]\!]$ les quatre conditions suivantes sont vérifiées.

1.
$$\exists\, a_i, b_i \in \mathbb{R} \quad \text{tels que} \quad a_i < b_i \quad \text{et} \quad P\big(Y_i \in [a_i\,;\,b_i]\big) \simeq 1. \tag{A.11}$$

2. La ddp de Y_i est continue et le pas de quantification assez petit :
$$\forall\, k \in [\![1\,;\,K_i - 2]\!],\, \exists\, p_{i,k} > 0 \quad \text{tel que} \quad \forall y_i \in]\eta_{i,k}\,;\,\eta_{i,k+1}],\;\; p(y_i) \simeq p_{i,k}.$$

3. Il existe une fonction strictement positive $\lambda_i \;:]a_i\,;\,b_i[\to \mathbb{R}_+^*$, continue, dont l'intégrale vaut 1, appelée densité de points de Qs_i, telle que pour tout $y \in]a_i\,;\,b_i[$, pour tout intervalle autour de y de longueur Δy suffisamment petite, le nombre de niveaux de quantification situés dans l'intervalle en question vaut approximativement $K_i \lambda_i(y) \Delta y$. On pose $\eta_{i,0} = a_i$ et $\eta_{i,K_i} = b_i$.

4. $\forall k \in [\![1\,;\,K_i]\!]$, $\hat{z}_{i,k} = \dfrac{\eta_{i,k} + \eta_{i,k-1}}{2}$.

A.4 Entropies, entropies jointes et information mutuelle

Dans le cas d'une quantification scalaire, notons $\mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_N$ l'ensemble des valeurs possibles du vecteur quantifié $\hat{\mathbf{Y}}$. Pour $\hat{\mathbf{y}}, \hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_V \in \mathcal{Z}$, nous posons

$$p(\hat{\mathbf{y}}) = P(\hat{\mathbf{Y}} = \hat{\mathbf{y}}) \quad \text{et} \quad p(\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_V) = P\Big(\hat{\mathbf{Y}}^V = (\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_V)\Big).$$

De même pour $i \in [\![1\,;\,N]\!]$ et $\hat{y}_i, \hat{y}_{i,1}, \ldots, \hat{y}_{i,V} \in \mathcal{Z}_i$, nous posons

$$p(\hat{y}_i) = P(\hat{Y}_i = \hat{y}_i), \qquad p(\hat{y}_{i,1}, \ldots, \hat{y}_{i,V}) = P\Big(\hat{\mathbf{Y}}_i^V = (\hat{y}_{i,1}, \ldots, \hat{y}_{i,V})\Big),$$

L'entropie du vecteur $\hat{\mathbf{Y}}$, exprimée en bits, est définie par

$$H(\hat{\mathbf{Y}}) = -\,\mathrm{E}[\log_2[p(\hat{\mathbf{Y}})] = -\sum_{\hat{\mathbf{y}} \in \mathcal{Z}} p(\hat{\mathbf{y}}) \log_2[p(\hat{\mathbf{y}})]$$

et celle de la séquence $\hat{\mathbf{Y}}^V$ par

$$H(\hat{\mathbf{Y}}^V) = -\sum_{(\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_V) \in \mathcal{Z}^V} p(\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_V) \log_2[p(\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_V)].$$

Dans le cas d'une quantification vectorielle de dimension V par composante, notons $\mathcal{Z}_{i,V}$ l'ensemble des valeurs possibles de la $i^{\text{ème}}$ composante quantifiée $\hat{\mathbf{Y}}_i^V$ et $\mathcal{Z}_V = \mathcal{Z}_{1,V} \times \cdots \times \mathcal{Z}_{N,V}$ l'ensemble des valeurs possibles de l'image transformée quantifiée $\hat{\mathbf{Y}}^V$. Pour $\hat{\mathbf{y}}_i^V \in \mathcal{Z}_{i,V}$ et $\hat{\mathbf{y}}^V \in \mathcal{Z}_V$ nous posons

$$p(\hat{\mathbf{y}}_i^V) = P(\hat{\mathbf{Y}}_i^V = \hat{\mathbf{y}}_i^V) \quad \text{et} \quad p(\hat{\mathbf{y}}^V) = P(\hat{\mathbf{Y}}^V = \hat{\mathbf{y}}^V).$$

L'entropie de $\hat{\mathbf{Y}}_i^V$ vaut

$$H(\hat{\mathbf{Y}}_i^V) = -\sum_{\hat{\mathbf{y}}_i^V \in \mathcal{Z}_{i,V}} p(\hat{\mathbf{y}}_i^V) \log_2[p(\hat{\mathbf{y}}_i^V)]$$

[4]Voir le paragraphe 5.6, pp. 161–167 du livre de A. Gersho et R. M. Gray, *Vector quantization and signal compression*, Kluwer, 1992.

et celle de $\hat{\mathbf{Y}}^V$ vaut :

$$H(\hat{\mathbf{Y}}^V) = -\sum_{\hat{\mathbf{y}}^V \in \mathcal{Z}_V} p(\hat{\mathbf{y}}^V) \log_2[p(\hat{\mathbf{y}}^V)].$$

De même, on introduit l'entropie du vecteur aléatoire continu, encore appelée entropie différentielle de $\mathbf{Y}$,

$$h(\mathbf{Y}) = -\,\mathrm{E}[\log_2[p(\mathbf{Y})]] = -\int_{\mathbb{R}^N} p(\mathbf{y}) \log_2[p(\mathbf{y})]\, d\mathbf{y}$$

et l'entropie jointe

$$h(\mathbf{Y}_1, \ldots, \mathbf{Y}_V) = -\int_{\mathbb{R}^N} \cdots \int_{\mathbb{R}^N} p(\mathbf{y}_1, \ldots, \mathbf{y}_V) \log_2[p(\mathbf{y}_1, \ldots, \mathbf{y}_V)]\, d\mathbf{y}_1 \cdots d\mathbf{y}_V.$$

Quand pour chaque composante la quantification est vectorielle en réseau et à haute résolution, l'entropie de l'image hyperspectrale quantifiée $H(\hat{\mathbf{Y}}^V)$ peut être reliée à l'entropie différentielle $h(\mathbf{Y}^V)$ de l'image hyperspectrale non quantifiée[5] :

$$H(\hat{\mathbf{Y}}_i^V) \simeq h(\mathbf{Y}_i^V) - \log_2[\delta_i^V |\det \mathbf{U}_i|]. \tag{A.12}$$

Quand pour chaque composante la quantification est scalaire à haute résolution, l'entropie de l'image hyperspectrale quantifiée $H(\hat{\mathbf{Y}}^V)$ peut être reliée à l'entropie différentielle $h(\mathbf{Y}^V)$ de l'image

[5]En effet,

$$\begin{aligned} H(\hat{\mathbf{Y}}_i^V) &= -\sum_{k=1}^{K_i} P\left(\mathbf{Y}_i^V \in \mathrm{cel}(k)\right) \log_2\left[P\left(\mathbf{Y}_i^V \in \mathrm{cel}(k)\right)\right] \\ &\simeq -\sum_{k=1}^{K_i} p_{i,k}\, \mathrm{vol}(k) \log_2[p_{i,k}\, \mathrm{vol}(k)] \simeq h(\mathbf{Y}_i^V) - \log_2[\delta_i^V |\det \mathbf{U}_i|]. \end{aligned}$$

hyperspectrale non quantifiée[6] :

$$H(\hat{\mathbf{Y}}^V) \simeq h(\mathbf{Y}^V) - \sum_{\ell=1}^{V}\sum_{i=1}^{N} \mathrm{E}\left[\log_2\left(\frac{1}{K_i\lambda_i(Y_{i,\ell})}\right)\right]. \tag{A.13}$$

Dans le cas où pour chaque composante la quantification est uniforme, de pas de quantification

[6]En effet, posons $\mathcal{K} = [\![1\,;\,K_1]\!] \times [\![1\,;\,K_2]\!] \times \cdots \times [\![1\,;\,K_N]\!]$, pour $\mathbf{k} \in \mathcal{K}$, $\mathbf{k} = (k_1, \ldots, k_N)^T$, notons $\mathrm{cel}(\mathbf{k})$ le pavé de $\mathbb{R}^N$ défini par

$$\mathrm{cel}(\mathbf{k}) = [\eta_{1,k_1-1}\,;\,\eta_{1,k_1}] \times \cdots \times [\eta_{N,k_N-1}\,;\,\eta_{N,k_N}]$$

et $\mathrm{vol}(\mathbf{k})$ son volume :

$$\mathrm{vol}(\mathbf{k}) = \prod_{i=1}^{N}(\eta_{i,k_i} - \eta_{i,k_i-1}).$$

à tout V-uplet $(\mathbf{k}_1, \ldots, \mathbf{k}_V) \in \mathcal{K}^V$, associons, grâce au théorème de la valeur moyenne, $\mathbf{z}(\mathbf{k}_1)$, …, $\mathbf{z}(\mathbf{k}_T) \in \mathbb{R}^N$ tel que $\forall \ell \in [\![1\,;\,V]\!]$, $\mathbf{z}(\mathbf{k}_\ell) \in \mathrm{cel}(\mathbf{k}_\ell)$ et

$$P\Big(\mathbf{Y}_1 \in \mathrm{cel}(\mathbf{k}_1) \text{ et } \cdots \text{ et } \mathbf{Y}_V \in \mathrm{cel}(\mathbf{k}_V)\Big) = p_{\mathbf{Y}^V}[\mathbf{z}(\mathbf{k}_1), \ldots, \mathbf{z}(\mathbf{k}_V)] \prod_{\ell=1}^{V} \mathrm{vol}(\mathbf{k}_\ell),$$

où $p_{\mathbf{X}^T}(\mathbf{x}^T)$ est la ddp jointe (supposée continue) des T vecteurs aléatoires $(\mathbf{X}_1, \ldots, \mathbf{X}_T)$ au point $\mathbf{x}^T \in (\mathbb{R}^N)^T$. Pour des quantifications scalaires à haute résolution, on a, pour $\ell \in [\![1\,;\,T]\!]$, pour $\mathbf{k}_\ell = (k_{1\ell}, \ldots, k_{N\ell})^T \in \mathcal{K}$, en notant $z_i(\mathbf{k}_\ell)$ la $i^{\text{ème}}$ composante de $\mathbf{z}(\mathbf{k}_\ell)$

$$z_i(\mathbf{k}_\ell) \in [\eta_{i,k_{i\ell}-1}\,,\,\eta_{i,k_{i\ell}}] \quad \text{ et } \quad \eta_{i,k_{i,\ell}} - \eta_{i,k_{i,\ell}-1} \simeq \frac{1}{K_i\lambda_i[z_i(\mathbf{k}_\ell)]},$$

donc

$$\mathrm{vol}(\mathbf{k}_\ell) \simeq \prod_{i=1}^{N} \frac{1}{K_i\lambda_i[z_i(\mathbf{k}_\ell)]}.$$

Par définition de l'entropie de $\hat{\mathbf{Y}}^V$, on a

$$\begin{aligned}
H(\hat{\mathbf{Y}}^V) \;&=\; -\sum_{\mathbf{k}_1\in\mathcal{K}} \cdots \sum_{\mathbf{k}_V\in\mathcal{K}} p_{\mathbf{Y}^V}[\mathbf{z}(\mathbf{k}_1), \ldots, \mathbf{z}(\mathbf{k}_V)] \left(\prod_{\ell=1}^{V} \mathrm{vol}(\mathbf{k}_\ell)\right) \log_2\left[p_{\mathbf{Y}^V}[\mathbf{z}(\mathbf{k}_1), \ldots, \mathbf{z}(\mathbf{k}_V)] \prod_{\ell=1}^{V} \mathrm{vol}(\mathbf{k}_\ell)\right] \\
&\simeq\; -\int_{(\mathbb{R}^N)^V} p_{\mathbf{Y}^V}(\mathbf{y}^V) \log_2[p_{\mathbf{Y}^V}(\mathbf{y}^V)]\, d\mathbf{y}^V \\
&\qquad -\sum_{\mathbf{k}_1\in\mathcal{K}} \cdots \sum_{\mathbf{k}_V\in\mathcal{K}} p_{\mathbf{Y}^V}[\mathbf{z}(\mathbf{k}_1), \ldots, \mathbf{z}(\mathbf{k}_V)] \log_2\left[\prod_{\ell=1}^{V} \mathrm{vol}(\mathbf{k}_\ell)\right] \left(\prod_{\ell=1}^{V} \mathrm{vol}(\mathbf{k}_\ell)\right) \\
&\simeq\; h(\mathbf{Y}^V) - \mathrm{E}\left[\log_2\left(\prod_{\ell=1}^{V}\prod_{i=1}^{N} \frac{1}{K_i\lambda_i(Y_{i,\ell})}\right)\right],
\end{aligned}$$

ce qui donne la relation (A.13). Pour $V = 1$ et $N = 1$, on retrouve la relation (9.9.6), page 298 du livre de Gersho et Gray dont les références sont données à la note du bas de la page 72.

$\delta_i = \eta_{i,k} - \eta_{i,k-1}$ $(1 \leq i \leq N)$ ces expressions deviennent[7]

$$H(\hat{\mathbf{Y}}) \simeq h(\mathbf{Y}) - \sum_{i=1}^{N} \log_2(\delta_i) \tag{A.14}$$

$$H(\hat{\mathbf{Y}}^V) \simeq h(\mathbf{Y}^V) - V \sum_{i=1}^{N} \log_2(\delta_i). \tag{A.15}$$

Pour un processus aléatoire stationnaire au sens strict $(\mathbf{Y}_k)_{k\in\mathbb{Z}}$, qui après quantification scalaire donne la suite $(\hat{\mathbf{Y}}_k)_{k\in\mathbb{Z}}$ ($\hat{\mathbf{Y}}_k = \mathrm{Qs}[\mathbf{Y}_k]$), nous appelons *entropie d'ordre* V de $(\hat{\mathbf{Y}}_k)_{k\in\mathbb{Z}}$ l'entropie jointe de $\hat{\mathbf{Y}}^V$, en particulier pour $V = 1$, l'entropie $H(\hat{\mathbf{Y}})$ de $\hat{\mathbf{Y}}_k$ est appelée *entropie du premier ordre* de $(\hat{\mathbf{Y}}_k)_{k\in\mathbb{Z}}$.

Par ailleurs, par définition de l'information mutuelle $I(\hat{\mathbf{Y}}_1^V; \ldots; \hat{\mathbf{Y}}_N^V)$ entre les séquences $\hat{\mathbf{Y}}_1^V$, …, $\hat{\mathbf{Y}}_N^V$, il vient

$$H(\hat{\mathbf{Y}}^V) = \sum_{i=1}^{N} H(\hat{\mathbf{Y}}_i^V) - I(\hat{\mathbf{Y}}_1^V; \ldots; \hat{\mathbf{Y}}_N^V).$$

Nous avons vu que le gain de codage peut s'exprimer en fonction des entropies $H(\hat{\mathbf{Y}}_i^V)$, et pour V grand il paraît plus fiable d'estimer le débit d'entropie plutôt que l'entropie d'ordre V. C'est pourquoi nous présentons la notion de débit d'entropie.

A.4.1 Débit d'entropie

Quand le processus $\hat{\mathbf{Y}}_1$, $\hat{\mathbf{Y}}_2$, …, $\hat{\mathbf{Y}}_V$, …, est stationnaire au sens strict, le rapport $H(\hat{\mathbf{Y}}^V)/V$ admet une limite quand V tend vers l'infini :

$$\lim_{V\to+\infty} \frac{H(\hat{\mathbf{Y}}^V)}{V} = H(\hat{\mathcal{Y}}).$$

La limite est appelée *débit entropique* (*entropy rate*) du processus $(\hat{\mathbf{Y}}_k)_{k\in\mathbb{Z}}$. Un théorème de la théorie de l'information[8] affirme que dans le cas d'un processus $(\hat{\mathbf{Y}}_k)_{k\in\mathbb{Z}}$ stationnaire, la suite des entropies conditionnelles $\big(H(\hat{\mathbf{Y}}_V \,|\, \hat{\mathbf{Y}}_{V-1}, \ldots, \hat{\mathbf{Y}}_1)\big)_{V\in\mathbb{N}}$, où

$$H(\hat{\mathbf{Y}}_V \,|\, \hat{\mathbf{Y}}_{V-1}, \ldots, \hat{\mathbf{Y}}_1) = -\,\mathrm{E}\left[\log_2\left(\frac{p(\hat{\mathbf{Y}}_1, \ldots, \hat{\mathbf{Y}}_V)}{p(\hat{\mathbf{Y}}_1, \ldots, \hat{\mathbf{Y}}_{V-1})}\right)\right],$$

est décroissante et tend vers le débit entropique $H(\hat{\mathcal{Y}})$ quand V tend vers l'infini.

De même, quand le processus aléatoire continu $(\mathbf{Y}_k)_{k\in\mathbb{Z}}$ est stationnaire au sens strict, le rapport $h(\mathbf{Y}^V)/V$ converge[9] quand V tend vers l'infini, sa limite $h(\mathcal{Y})$ est le *débit entropique* du processus continu $(\mathbf{Y}_k)_{k\in\mathbb{Z}}$ et d'après la relation (A.14), dans le cas d'une quantification scalaire uniforme par composante, on a

$$H(\hat{\mathcal{Y}}) \simeq h(\mathcal{Y}) - \sum_{i=1}^{N} \log_2 \delta_i.$$

[7] On retrouve ainsi le théorème 9.3.1 page 229 du livre de Cover et Thomas (voir la note page 69), énoncé pour $N = 1$ et $V = 1$.

[8] Ibid., théorèmes 4.2.1 et 4.2.2 page 64.

[9] Ibid., paragraphe 11.5 page 273.

De la même façon, en ne considérant qu'une composante (la $i^{\text{ème}}$) du processus $(\mathbf{Y}_k)_k$, en supposant la quantification uniforme sur cette composante et en introduisant les débits entropiques respectifs $H(\hat{\mathcal{Y}}_i)$ et $h(\mathcal{Y}_i)$ des processus $(\hat{Y}_{i,k})_{k\in\mathbb{Z}}$ et $(Y_{i,k})_{k\in\mathbb{Z}}$, on a

$$H(\hat{\mathcal{Y}}_i) \simeq h(\mathcal{Y}_i) - \log_2 \delta_i.$$

A.5 Distorsion d'un quantificateur à haute résolution

A.5.1 Cas d'un quantificateur vectoriel

Pour un coût quadratique, la distorsion moyenne par échantillon due à la quantification vectorielle de la $i^{\text{ème}}$ composante s'écrit :

$$D_i = \frac{1}{V}\,\mathrm{E}\Big[\big\|\mathbf{Y}_i^V - \mathrm{Qv}_i[\mathbf{Y}_i^V]\big\|^2\Big].$$

On montre que sous l'hypothèse d'une quantification vectorielle en réseau à haute résolution, la distorsion D_i se met sous la forme[10] :

$$D_i \simeq \frac{\delta_i^2 M(\mathbf{U}_i)}{V|\det \mathbf{U}_i|}. \tag{A.16}$$

A.5.2 Cas d'un quantificateur scalaire

Pour un coût quadratique, la distorsion due à la quantification de la $i^{\text{ème}}$ composante s'écrit :

$$D_i = \mathrm{E}\Big[\big|X_i - \mathrm{Qs}_i[X_i]\big|^2\Big].$$

On montre que sous l'hypothèse d'une quantification à haute résolution, la distorsion D_i se met sous la forme[11] :

$$D_i \simeq \frac{1}{12K_i^2}\int_{a_i}^{b_i} \frac{p(y_i)}{\lambda_i^2(y_i)}\,dy_i, \tag{A.17}$$

[10] En effet, comme dans le livre de Gersho et Gray (vor les références à la note 4), la distorsion D_i due à Qv_i vérifie $VD_i = \sum_{k=1}^{K_i}\int_{\mathrm{cel}(k)} \big\|\mathbf{y}_i^V - \hat{\mathbf{z}}_i^V(k)\big\|^2 p(\mathbf{y}_i^V)\,d\mathbf{y}_i^V \simeq \sum_{k=1}^{K_i} p_{i,k}\int_{\mathrm{cel}(k)} \big\|\mathbf{y}^V - \hat{\mathbf{z}}_i^V(k)\big\|^2 d\mathbf{y}^V = \frac{\delta_i^2 M(\mathbf{U}_i)}{|\det \mathbf{U}_i|}\sum_{k=1}^{K_i} p_{i,k}\,\mathrm{vol}(k) \simeq \frac{\delta_i^2 M(\mathbf{U}_i)}{|\det \mathbf{U}_i|}$.

[11] En effet, comme cela est fait dans le livre de Gersho et Gray page 163, la distorsion due à Qs_i vaut :

$$D_i = \sum_{k=1}^{K_i}\int_{\eta_{i,k-1}}^{\eta_{i,k}} (y_i - \hat{z}_{i,k})^2 p(y_i)\,dy_i \simeq \sum_{k=1}^{K_i} p_{i,k}\int_{\eta_{i,k-1}}^{\eta_{i,k}} (y_i - \hat{z}_{i,k})^2\,dy_i = \sum_{k=1}^{K_i}\frac{p_{i,k}(\eta_{i,k}-\eta_{i,k-1})^3}{12}.$$

Or $P(y_i = \hat{z}_{i,k}) = \displaystyle\int_{\eta_{i,k-1}}^{\eta_{i,k}} p(y_i)dy_i \simeq p_{i,k}(\eta_{i,k}-\eta_{i,k-1})$, donc $D_i \simeq \dfrac{1}{12}\displaystyle\sum_{k=1}^{K_i} P(Y_i = \hat{z}_{i,k})(\eta_{i,k}-\eta_{i,k-1})^2$. Enfin, il résulte de la condition 3. de la quantification à haute résolution, l'intervalle $]\eta_{i,k-1}\,;\,\eta_{i,k}]$ ne contenant qu'une cellule de Qs_i, que $\eta_{i,k}-\eta_{i,k-1} \simeq \dfrac{1}{K_i\lambda_i(\hat{z}_{i,k})}$, cela entraîne avec la relation précédente et la suivante : $P(Y_i = \hat{z}_{i,k}) \simeq p_{Y_i}(\hat{z}_{i,k})(\eta_{i,k}-\eta_{i,k-1})$, où $p_{Y_i}(\hat{z}_{i,k})$ désigne la ddp de Y_i au point $y_i = \hat{z}_{i,k}$, que

$$D_i \simeq \frac{1}{12}\sum_{k=1}^{K_i} P(Y_i = \hat{z}_{i,k})\big(K_i\lambda_i(\hat{z}_{i,k})\big)^{-2} \simeq \frac{1}{12}\sum_{k=1}^{K_i} p_{Y_i}(\hat{z}_{i,k})(\eta_{i,k}-\eta_{i,k-1})\big(K_i\lambda_i(\hat{z}_{i,k})\big)^{-2}.$$

D'après la continuité de $p(y_i)$ et $\lambda_i(y_i)$, cette dernière expression est équivalente à une intégrale, donnant ainsi la relation (A.17).

on obtient ainsi la formule de Bennett qui s'écrit aussi

$$D_i \simeq \frac{1}{12}\,\mathrm{E}\left[\frac{1}{K_i^2\lambda_i(Y_i)^2}\right]. \tag{A.18}$$

A.6 Distorsion et entropie d'ordre V

A.6.1 Cas de la quantification vectorielle en réseau

La quantification vectorielle Qv_i est appliquée à V échantillons de la $i^{\text{ème}}$ composante. Sous les hypothèses d'une quantification en réseau et à haute résolution, nous déduisons des relations (A.12) et (A.16) que le débit $R_i^{\mathrm{Qv}}(D_i)$ exprimé en bits par échantillon vaut

$$\begin{aligned} R_i^{\mathrm{Qv}}(D_i) &= \frac{H(\hat{\mathbf{Y}}_i^V)}{V} \simeq \frac{h(\mathbf{Y}_i^V)}{V} - \frac{1}{2}\log_2\left[\delta_i^2|\det \mathbf{U}_i|^{1/V}\right] \\ &\simeq \frac{h(\mathbf{Y}_i^V)}{V} - \frac{1}{2}\log_2\left[\frac{D_i}{G(\mathbf{U}_i)}\right] \end{aligned}$$

où la distorsion D_i est une distorsion moyenne par échantillon et où la constante $G(\mathbf{U}_i)$ ne dépend que de la forme du réseau de Qv_i (i.e., de $\mathbf{U}_i$) :

$$G(\mathbf{U}_i) = \frac{M(\mathbf{U}_i)}{V|\det \mathbf{U}_i|^{1+2/V}},$$

en particulier pour $V = 1$ on trouve le facteur $G(1) = 1/12$.

A.6.2 Cas de la quantification scalaire

La quantification scalaire Qs_i est appliquée à V échantillons $(Y_{i,\ell})_{1\le\ell\le V}$ de la $i^{\text{ème}}$ composante. Sous les hypothèses d'une quantification à haute résolution, nous déduisons de la relation (A.13) que

$$H(\hat{\mathbf{Y}}_i^V) \simeq h(\mathbf{Y}_i^V) - \frac{1}{2}\sum_{\ell=1}^{V}\mathrm{E}\left(\log_2\left[\frac{1}{K_i^2\lambda_i(Y_{i,\ell})^2}\right]\right).$$

L'inégalité de Jensen et la relation (A.17) entraînent

$$-\,\mathrm{E}\left(\log_2\left[\frac{1}{K_i^2\lambda_i(Y_{i,\ell})^2}\right]\right) \ge -\log_2\left[\mathrm{E}\left(\frac{1}{K_i^2\lambda_i(Y_{i,\ell})^2}\right)\right] \simeq -\log_2[12D_{i,\ell}],$$

avec égalité (dans l'inégalité de Jensen) si et seulement si la densité de point λ_i est constante $\lambda_i(y_i) = \frac{1}{b_i-a_i}$ (i.e., la quantification Qs_i est uniforme), et dans ce cas la distorsion $D_{i,\ell}$ ne dépend pas de ℓ :

$$D_{i,\ell} \simeq \frac{1}{12K_i^2(b_i-a_i)^2} \simeq D_i.$$

En conclusion, il est établi que pour une distorsion moyenne $D_i = \frac{1}{V}\sum_{\ell=1}^{V} D_{i,\ell}$ donnée, l'entropie jointe de V valeurs quantifiées est minimale si et seulement si la quantification est uniforme et l'entropie minimale, exprimée en bits par échantillon, vaut alors

$$R_{i,\mathrm{opt}}^{\mathrm{Qs}}(D_i) = \frac{H(\hat{\mathbf{Y}}_i^V)}{V} \simeq \frac{h(\mathbf{Y}_i^V)}{V} - \frac{1}{2}\log_2[12D_i]. \tag{A.19}$$

La cas $V = 1$ est traité dans le livre de Gersho et Gray pp. 297–300.

Inversement, pour une entropie jointe de V valeurs quantifiées $H(\hat{\mathbf{Y}}_i^V) = VR_i$ donnée, quel est le minimum de la distorsion moyenne $D_i = \frac{1}{V}\sum_{\ell=1}^{V} D_{i,\ell}$, où $D_{i,\ell} = \mathrm{E}\left[|Y_{i,\ell} - \hat{Y}_{i,\ell}|^2\right]$, sous l'hypothèse d'une quantification à haute résolution ?

Le théorème de la moyenne nous dit que

$$\frac{12}{V}\sum_{\ell=1}^{V} D_{i,\ell} \geq 12\left(\prod_{\ell=1}^{V} D_{i,\ell}\right)^{1/V} \tag{A.20}$$

avec égalité si et seulement si les distorsions $D_{i,\ell}$ $(1 \leq \ell \leq V)$ sont identiques, égales à $D_i > 0$. Par ailleurs on a

$$\log_2\left[12\left(\prod_{\ell=1}^{V} D_{i,\ell}\right)^{1/V}\right] = \frac{1}{V}\sum_{\ell=1}^{V}\log_2[12D_{i,\ell}] \simeq \frac{1}{V}\sum_{\ell=1}^{V}\log_2\left[\mathrm{E}\left(\frac{1}{K_i^2\lambda_i^2(Y_{i,\ell})}\right)\right]$$

d'après la formule de Bennet ; et l'inégalité de Jensen donne

$$\frac{1}{V}\sum_{\ell=1}^{V}\log_2\left[\mathrm{E}\left(\frac{1}{K_i^2\lambda_i^2(Y_{i,\ell})}\right)\right] \geq \frac{2}{V}\sum_{\ell=1}^{V}\mathrm{E}\left(\log_2\left[\frac{1}{K_i\lambda_i(Y_{i,\ell})}\right]\right)$$

avec égalité si et seulement si les quantifications Qs_i sont uniformes pour $1 \leq i \leq N$, enfin d'après la relation (A.13) (pour $N = 1$) on a

$$\frac{2}{V}\sum_{\ell=1}^{V}\mathrm{E}\left[\log_2\left(\frac{1}{K_i\lambda_i(Y_{i,\ell})}\right)\right] \simeq \frac{2}{V}h(\mathbf{Y}_i^V) - \frac{2}{V}H(\hat{\mathbf{Y}}_i^V).$$

Ainsi l'inégalité

$$\log_2\left[12\left(\prod_{\ell=1}^{V} D_{i,\ell}\right)^{1/V}\right] \geq \frac{2}{V}h(\mathbf{Y}_i^V) - \frac{2}{V}H(\hat{\mathbf{Y}}_i^V)$$

est toujours satisfaite, avec équivalence

$$\log_2\left[12\left(\prod_{\ell=1}^{V} D_{i,\ell}\right)^{1/V}\right] \simeq \frac{2}{V}h(\mathbf{X}_i^V) - \frac{2}{V}H(\hat{\mathbf{Y}}_i^V)$$

si et seulement si la quantification est uniforme. Dans ce cas on a égalité dans l'inégalité de la moyenne (A.20), ce qui entraîne

$$\log_2[12D_{i,\mathrm{opt}}^{\mathrm{Qs}}] \simeq \frac{2}{V}h(\mathbf{Y}_i^V) - \frac{2}{V}H(\hat{\mathbf{Y}}_i^V).$$

En conclusion on trouve que la distorsion est minimale si et seulement si la quantification est uniforme et dans ce cas elle vaut

$$D_{i,\mathrm{opt}}^{\mathrm{Qs}}(R_i) \simeq \frac{1}{12}2^{-2[R_i - h(\mathbf{Y}_i^V)/V]}. \tag{A.21}$$

A.7 Codage entropique d'ordre V

Dans ce paragraphe, nous noterons $\hat{\mathcal{Z}}_V$ l'ensemble des valeurs possibles du signal quantifié[12] $\hat{\mathbf{Y}}^V$ et $\hat{\mathcal{Z}}_{i,V}$ l'ensemble des valeurs possibles de $\hat{\mathbf{Y}}_i^V$, que la quantification soit vectorielle ou scalaire. Un *code source* $\mathcal{C}_V$ pour le signal quantifié $\hat{\mathbf{Y}}^V$ est une application de $\hat{\mathcal{Z}}_V$ dans $\{0,1\}^\star$, l'ensemble des suites finies de bits. Pour un élément $\hat{\mathbf{y}}^V \in \hat{\mathcal{Z}}_V$, on note $\mathcal{C}_V(\hat{\mathbf{y}}^V)$ le mot-code correspondant à $\hat{\mathbf{y}}^V$ et $\ell_{\mathcal{C}_V}(\hat{\mathbf{y}}^V)$ la longueur de $\mathcal{C}_V(\hat{\mathbf{y}}^V)$ exprimée en bits. La longueur moyenne $L(\mathcal{C}_V)$ du code source est définie par

$$L(\mathcal{C}_V) = \frac{1}{NV}\,\mathrm{E}[\ell_{\mathcal{C}_V}(\hat{\mathbf{Y}}^V)] = \frac{1}{NV}\sum_{\hat{\mathbf{y}}^V \in \hat{\mathcal{Z}}_V} p(\hat{\mathbf{y}}^V)\ell_{\mathcal{C}_V}(\hat{\mathbf{y}}^V),$$

elle est exprimée en bits par échantillon.

Un *code source* $\mathcal{C}_V$ est dit *régulier* (*non singular*) quand l'application $\mathcal{C}_V$ est injective. L'extension $\mathcal{C}_V^\star$ du code source $\mathcal{C}_V$ est l'application de $(\hat{\mathcal{Z}}_V)^\star$ — ensemble des séquences finies d'éléments de $\hat{\mathcal{Z}}_V$ — dans $\{0,1\}^\star$ définie par

$$\mathcal{C}_V^\star(\hat{\mathbf{y}}_1^V, \ldots, \hat{\mathbf{y}}_n^V) = \mathcal{C}_V(\hat{\mathbf{y}}_1^V)\mathcal{C}_V(\hat{\mathbf{y}}_2^V)\cdots\mathcal{C}_V(\hat{\mathbf{y}}_n^V),$$

où $\mathcal{C}_V(\hat{\mathbf{y}}_1^V)\mathcal{C}_V(\hat{\mathbf{y}}_2^V)\cdots\mathcal{C}_V(\hat{\mathbf{y}}_n^V)$ indique la concaténation des mots-codes correspondants. Un *code source* $\mathcal{C}_V$ est dit *déchiffrable* (*uniquely decodable*) quand son extension est régulière.

Un théorème de Shannon[13] affirme que

$$L(\mathcal{C}_V) \geq \frac{H(\hat{\mathbf{Y}}^V)}{NV}$$

pour n'importe quel code source $\mathcal{C}_V$ déchiffrable et un autre théorème de la théorie de l'information[14] affirme que le codeur de Huffman $\mathcal{C}_{V,\mathrm{Huf}}$ donne une longueur moyenne optimale (exprimée en bits par échantillon) vérifiant

$$\frac{H(\hat{\mathbf{Y}}^V)}{NV} \leq L(\mathcal{C}_{V,\mathrm{Huf}}) < \frac{H(\hat{\mathbf{Y}}^V)+1}{NV}. \tag{A.22}$$

Supposons maintenant que les composantes du signal vectoriel quantifié $\hat{\mathbf{Y}}^V$ soient codées séparément, c'est-à-dire indépendamment les unes des autres. Soit $\mathcal{C}_{iV} : \hat{\mathcal{Z}}_{i,V} \to \{0\,,\,1\}^\star$ le code source appliqué à la $i^{\text{ème}}$ composante, nous le supposons déchiffrable. À partir des N codes sources $\mathcal{C}_{iV}$ et d'une fonction de multiplexage arbitraire $f : \{0\,,\,1\}^\star \times \cdots \times \{0\,,\,1\}^\star \to \{0\,,\,1\}^\star$ mélangeant N flots de bits, définissons un code source $\mathcal{C}_V^s$ (que nous dirons *séparable*) :

$$\mathcal{C}_V^s(\hat{\mathbf{y}}^V) = f\left(\mathcal{C}_{1V}(\hat{\mathbf{y}}_1^V), \ldots, \mathcal{C}_{NV}(\hat{\mathbf{y}}_N^V)\right) \quad \text{pour tout} \quad \hat{\mathbf{y}}^V = \begin{pmatrix} \hat{\mathbf{y}}_1^V \\ \vdots \\ \hat{\mathbf{y}}_N^V \end{pmatrix}. \tag{A.23}$$

Sous cette condition, la longueur moyenne de code minimale $L_{V,\mathrm{opt}}^s$, exprimée en bits par échantillon, pour coder le signal quantifié $\hat{\mathbf{y}}^V$ vérifie

$$\frac{1}{V} + \frac{1}{NV}\sum_{i=1}^{N} H(\hat{\mathbf{Y}}_i^V) \geq L_{V,\mathrm{opt}}^s \geq \frac{1}{NV}\sum_{i=1}^{N} H(\hat{\mathbf{Y}}_i^V). \tag{A.24}$$

[12] Dans le cas d'une quantification vectorielle nous avons $\hat{\mathcal{Z}}_V = \mathcal{Z}_V$ et pour une quantification scalaire $\hat{\mathcal{Z}}_V = \mathcal{Z}^V$.

[13] Voir le théorème 5.3.1, page 86 du livre de Cover et Thomas décrit à la note de la page 69.

[14] Ibid., théorème 5.4.1 page 88 et théorème 5.8.1 page 101.

Annexe B

Articles récents parus

Signal Processing 90 (2010) 753–758

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Review

On optimal orthogonal transforms at high bit-rates using only second order statistics in multicomponent image coding with JPEG2000

Isidore Paul Akam Bita [a,*], Michel Barret [b], Dinh-Tuan Pham [c]

[a] Luxspace Sarl, Chateau de Betzdorf, L-6815 Betzdorf, Luxembourg
[b] Supelec, Information Multimodality and Signal Team, 2 rue É. Belin 57070 Metz, France
[c] Jean Kuntzmann Laboratory, 51 rue des Mathématiques, BP 53, 38041 Grenoble Cedex 9, France

ARTICLE INFO

Article history:
Received 17 October 2008
Received in revised form
28 April 2009
Accepted 24 August 2009
Available online 29 August 2009

Keywords:
Multicomponent image coding
High rate transform coding
Karhunen–Loève transform
Independent component analysis
JPEG2000
Multispectral image coding
Hyperspectral image coding

ABSTRACT

We study a JPEG2000 compatible multicomponent image compression scheme, which consists in applying a discrete wavelet transform (DWT) to each component of the image and a spectral linear transform between components. We consider the case of a spectral transform which adapts to the image and a 2-D DWT with fixed coefficients. In Akam Bita et al. (accepted for publication, [6]) we gave a criterion minimized by optimal spectral transforms. Here, we derive a simplified criterion by treating the transformed coefficients in each subband as having a Gaussian distribution of variance depending on the subband. Its minimization under orthogonality constraint is shown to lead to a joint approximate diagonalization problem, for which a fast algorithm (JADO) is available. Performances in coding of the transform returned by JADO are compared on hyper- and multi-spectral images with the Karhunen–Loève transform (KLT) and the optimal transform (without Gaussianity assumption) returned by the algorithm OrthOST introduced in Akam Bita et al. (accepted for publication, [6]). For hyper- (resp. multi-) spectral images, we observe that JADO returns a transform which performs appreciably better than (resp. as well as) the KLT at medium to high bit-rates, nearly attaining (resp. slightly below) the performances of the transform returned by OrthOST, with a significantly lower complexity than the algorithm OrthOST.

Contents

* Corresponding author. Tel.: +352 267 890 8046; fax: +352 267 890 4029.
E-mail addresses: AkamBita@luxspace.lu (I.P. Akam Bita), Michel.Barret@supelec.fr (M. Barret), Dinh-Tuan.Pham@imag.fr (D.-T. Pham).

0165-1684/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2009.08.008

1. Introduction

Research activities on multi- and hyper-spectral image compression have been intense these last years, because of its increasingly numerous applications (e.g., in satellite or aerial imaging). For multicomponent image compression, the JPEG2000 Part 2 standard [1] permits to apply either (1) a 3-D discrete wavelet transform (DWT) as in [2,3] or (2) only one linear spectral transform associated with 2-D DWT before quantization and entropy coding as in [4,5].

In [6] we introduced two algorithms that compute JPEG2000 compatible spectral transforms that are optimal in coding at high-bit rates, even when the data are not Gaussian. One, called `OrthOST`, returns an orthogonal optimal transform and the other, called `OST`, an optimal transform with no constraint but invertibility. We observed [6] that on hyperspectral images the orthogonal optimal spectral transform `OrthOST` performs slightly but significantly better than the Karhunen–Loève transform (KLT) at medium and high bit-rates—greater than 0.5 bits per pixel and per band (bpppb)—for four different measures of distortion, which give together a good idea of the codec performances on actual applications of hyperspectral images like classification and targets detections [7]. However, the drawback of this transform is its computational burden. In this paper we introduce a new algorithm that computes a quasi-optimal orthogonal spectral transform with a much smaller computational complexity than `OrthOST` and we test the performance of this transform. We shall use the following notations.

Let $\mathbf{X}$ be a multicomponent image with N components $\mathbf{X}_1,\ldots,\mathbf{X}_N$. Each component $\mathbf{X}_i$ is a 2-D image with N_r rows and N_c columns. In order to simplify the notations and the mathematical expressions, we assume that each component is written as a row vector by scanning all its pixels row by row. Then $\mathbf{X}$ is a $N\times L$ matrix, with $L=N_rN_c$. The description of a compression scheme compatible with the JPEG2000 Part 2 standard can be summarized as follows (we called it *separable scheme* in [6]):

- *Coding*. The same 2-D DWT is applied to each component $\mathbf{X}_i$. We denote $\mathbf{W}$ the invertible $L\times L$ matrix associated with that DWT. The result of the 2-D DWT applied to the entire image is $\mathbf{XW}^T$, where ^{T} denotes transposition. Further, a linear transform $\mathbf{A}$ is applied in order to reduce the spectral redundancies between the components. Hence, the transformed coefficients are the elements of the matrix $\mathbf{Y}=\mathbf{AXW}^T$, whose i th row is denoted $\mathbf{Y}_i$. Finally, the transformed coefficients are quantized, with a specific scalar quantizer per subband and per component, before entropy coded.
- *Decoding*. Let $\mathbf{Y}^q$ denote the matrix with the same dimension as $\mathbf{Y}$ containing the dequantized transformed coefficients. The mathematical inverse transforms are applied to $\mathbf{Y}^q$ in order to reconstruct an approximation $\hat{\mathbf{X}}=\mathbf{A}^{-1}\mathbf{Y}^q\mathbf{W}^{-T}$ of the original image $\mathbf{X}$, with the components $\hat{\mathbf{X}}_1,\ldots,\hat{\mathbf{X}}_N$.

In the next section we recall the criterion that gives the optimal spectral transforms [6], when minimized. This criterion involves differential entropies of the transformed coefficient components. Then, in Section 3 we justify the interest of a simplified criterion which involves only second order statistics. After deriving the modified criterion, we show that its minimization (under the orthogonality constraint) is equivalent to a joint approximate diagonalization, for which a fast algorithm is available [8]. A new algorithm, called `JADO`, that returns an orthogonal transforms minimizing the new criterion is presented in Appendix A. Finally, for multi- and hyperspectral images, the performances in coding of the different spectral transforms mentioned above are presented in Section 4.

2. Criterion minimized by optimal JPEG2000 spectral transforms at high bit-rates

Let M be the number of subbands per component after the 2-D DWT and π_m $(1\leq m\leq M)$ be the proportion, in a component, of transformed coefficients that belong to the subband m, therefore $\sum_{m=1}^{M}\pi_m=1$. The aim of the DWT is to reduce the spatial redundancies in each component. We are interested in searching spectral transforms that adapt to the data in order to be optimal in coding when the 2-D DWT has fixed coefficients. Actually in all our tests the 2-D DWT is the so-called Daubechies 9/7, recommended for lossy compression in JPEG2000 standard [1]. In [6, Theorem 4], we showed that under high resolution quantization hypotheses[1], an optimal spectral transform $\mathbf{A}$ (i.e. which minimizes the total bit-rate of the separable scheme, for a given end-to-end mean square error) minimizes the criterion

$$C(\mathbf{A})=\sum_{j=1}^{N}\sum_{m=1}^{M}\pi_m H(Y_j^{(m)})+\frac{1}{2}\log_2\det\operatorname{diag}(\mathbf{A}^{-T}\mathbf{A}^{-1}), \qquad (1)$$

where diag is the diagonal matrix obtained with the diagonal elements of its argument, det the determinant and $H(Y_j^{(m)})$ the differential entropy of the transformed coefficients that belong to the m th subband of the j th component of $\mathbf{Y}$. However, the algorithms `OrthOST` and `OST` developed in [6] to minimize this optimality criterion are costly in computations compared with the one that computes the usual KLT. Note that the algorithm that computes a KLT is customarily applied first to the image before the DWT, but this would be equivalent to applying it after the DWT (i.e. to the DWT coefficients) if the DWT is orthogonal (as is often the case or at least nearly so in practice[2]). But then it will not distinguish subbands: the

[1] These hypotheses, sometimes known as the asymptotic quantization approach (see [9] sections 5.6 and 9.9) can be stated as: (1) the random vector $\mathbf{Y}$ has a continuous probability density function f_Y; and (2) the quantization steps h of the N components are small with respect to the variations of f_Y (i.e. $f_Y(y+h)'\ f_Y(y)$, $8y\ 2\ R_N$); and (3) centroid condition: for any cell S of the separable N-D quantizer, the dequantized value Y_q associated with S satisfies $Y_q=E[Y|Y2S]$, where E denotes the mathematical expectation

[2] We have $(\mathbf{XW}^T)(\mathbf{XW}^T)^T\approx\mathbf{XX}^T$ when $\mathbf{W}^T\mathbf{W}\approx\mathrm{Id}$ (the identity matrix) as it is roughly the case with the approximately orthogonal Daubechies 9/7 DWT (indeed, a simulation shows that $\|\mathbf{W}^T\mathbf{W}-\mathrm{Id}\|_\infty=0.42$ and the infinity norm of the off diagonal of $\mathbf{W}^T\mathbf{W}$ is worth 0.16 for five levels of decomposition on a 1-D signal of length 512).

DWT coefficients are considered as coming from a same (Gaussian) distribution, regardless of the subband they belong to. We feel that the higher performance of criterion (1) over the criterion $\frac{1}{2}\log_2\prod_{j=1}^{N}\mathrm{var}(Y_j)$, which leads to the KLT, is due primarily to the fact that it treats each subband separately rather than that treating the distribution in each subband as non-Gaussian. This is logical since after any DWT, the energy in each subband depends on the power spectrum of the input signal. It is important to notice that there is no contradiction in the fact that the criterion (1) treats each subband separately, while the same spectral transform **A** is applied to all the subbands. The idea is then to introduce the distinction between subbands but retain the (approximate) Gaussian used by the KLT. The distribution of all the wavelet coefficients (with no distinction between subbands) is a mixture of distributions of the coefficients in the subbands. It can be shown that the kurtosis of the mixed distribution is higher than the average kurtosis of the individual distributions. In particular, mixture of Gaussian distributions has always a positive kurtosis, unless all the individual distributions are the same. Thus the wavelet coefficients, regardless of the subband they belong to, have a positive kurtosis even if in each subband their distribution is Gaussian. The above consideration suggests modifying the criterion (1) by treating the $Y_j^{(m)}$ in each subband m as having a Gaussian distribution *with differing variance for different m*. The transformation minimizing this modified criterion is no longer optimal, but can be nearly so if the distribution of each $Y_j^{(m)}$ is not too far from Gaussian. This is not an unrealistic situation: the wavelet coefficients in a subband is the (decimated) output of a bandpass filter which tends to produce more Gaussian output than input, due to the reasoning (given e.g. in [10] section 8–5) that yields to the proof of the Central Limit Theorem. The advantage of the modified criterion is that it avoids the entropy estimation and uses only second order statistics. Thus its minimization requires much less computer resources than using (1).

3. A simplified criterion using only second order statistics

Let $H^-(Z)=\log_2\sqrt{\mathrm{var}(Z)2\pi e}-H(Z)$ denote the negentropy of Z (which is the difference of entropy between a Gaussian distribution with variance $\mathrm{var}(Z)$ and the distribution of Z), it is non-negative and vanishes if and only if Z is Gaussian. The criterion (1) can be rewritten for orthogonal[3] matrices

$$C_\perp(\mathbf{A})=-\sum_{j=1}^{N}\sum_{m=1}^{M}\pi_m H^-(Y_j^{(m)}) + \frac{1}{2}\sum_{j=1}^{N}\sum_{m=1}^{M}\pi_m\log_2[\mathrm{var}(Y_j^{(m)})2\pi e]. \quad (2)$$

[3] The orthogonality constraint is justified by our earlier work [6] in which we found that minimizing (1) with and without this constraint yields almost the same performances. With the orthogonality constraint, the second term in (1) vanishes.

An analysis of criterion (2) shows that it takes into account two phenomena: (1) the non-Gaussianity of the transformed coefficients $Y_j^{(m)}$ for $1\le m\le M$ and $1\le j\le N$—this is controlled by the first term—and (2) the inhomogeneity of the variances in the subbands—this is controlled by the second term. It is natural to explore the case where the second phenomenon is the most important, since the DWT tends to render the variables more Gaussian. In practice, this condition is generally roughly satisfied, except in the LL subband (a subband of lowest resolution) for which the weighting coefficient π_m is generally small. Thus, if we neglect the variation, induced by the spectral transform **A**, of the first term in the right member of Eq. (2), and if we consider only orthogonal matrices **A**, then the optimal transform minimizes the new criterion

$$C'(\mathbf{A})=\frac{1}{2}\sum_{j=1}^{N}\sum_{m=1}^{M}\pi_m\log_2[\mathrm{var}(Y_j^{(m)})]. \quad (3)$$

Furthermore if we assume in each component the transformed coefficients have all the same variance, regardless of the subband they belong to, then the criterion (3) becomes $\frac{1}{2}\log_2[\prod_{j=1}^{N}\mathrm{var}(Y_j)]$, leading to the KLT.

In the following, we express criterion (3) in terms of the covariance matrices of the wavelets coefficients $\mathbf{XW}^T$ located in the same subband. We begin by introducing new notations. After the 2-D DWT, for each component, the wavelet coefficients are regrouped subband by subband according to a fixed scan that does not depend on the component, this partitioning can be written as $\mathbf{XW}^T=[(\mathbf{XW}^T)^{(1)}\ (\mathbf{XW}^T)^{(2)}\ \cdots\ (\mathbf{XW}^T)^{(M)}]$. Indeed, the re-ordering of the wavelet coefficients corresponds to the right multiplication of $\mathbf{XW}^T$ by a permutation matrix **P**. We can suppose without loss of generality that this permutation is the identity, otherwise we may replace **W** with **PW**. The matrix $(\mathbf{XW}^T)^{(m)}$ is of dimension $N\times\pi_m L$. Its columns can be considered as different realizations of a random vector of dimension N whose covariance matrix is denoted $\mathbf{C}^{(m)}$. Now, $\mathbf{Y}=\mathbf{AXW}^T$ can be written $\mathbf{Y}=[\mathbf{Y}^{(1)}\ \cdots\ \mathbf{Y}^{(M)}]$, where $\mathbf{Y}^{(m)}=(\mathbf{AXW}^T)^{(m)}$ is a matrix whose columns can also be considered as different realizations of a random vector having $\mathbf{AC}^{(m)}\mathbf{A}^T$ as covariance matrix. With these notations, we have $\prod_{j=1}^{N}\mathrm{var}(Y_j^{(m)})=\det\mathrm{diag}(\mathbf{AC}^{(m)}\mathbf{A}^T)$ and hence the new criterion becomes

$$C'(\mathbf{A})=\frac{1}{2}\sum_{m=1}^{M}\pi_m\log_2\det\mathrm{diag}(\mathbf{AC}^{(m)}\mathbf{A}^T) \quad (4)$$

to be minimized with respect to **A**, under the constraint that it is orthogonal.

The FG algorithm in [8] can be used to minimize the above criterion. We have developed a slightly different algorithm (called JADO) which is briefly described in Appendix A.

4. Experimental results

In all our experiments, the 2-D DWT used is the Daubechies 9/7 with five levels of decomposition. The

Fig. 1. Some images used in our tests. From left to right: Moffett, Jasper, Cuprite.

tests have been executed with the Verification Model[4] (VM9), developed by the JPEG2000 group, which uses the EBCOT (Embedded Block Coding Optimization Truncation) coder [1] with its post compression rate-distortion optimizer applied across multiple bands. The performances are evaluated by the bit-rates and the associated end-to-end distortions. For hyperspectral images, we considered four distortions. A first one is the mean square error (MSE) expressed in terms of the SNR (signal to noise ratio), $\mathrm{SNR} = 10\log_{10}\sigma^2/D$ where

$$D = \frac{1}{NL}\sum_{i=1}^{N}\sum_{n=1}^{L}|X_i(n) - \hat{X}_i(n)|^2$$

is the actual end-to-end MSE distortion and σ^2 is the empirical variance of the initial image:

$$\sigma^2 = \frac{1}{NL}\sum_{i=1}^{N}\sum_{n=1}^{L}(X_i(n) - \mu)^2,$$

with

$$\mu = \frac{1}{NL}\sum_{i=1}^{N}\sum_{n=1}^{L}X_i(n)$$

the empirical mean of the image. A second distortion is the maximum absolute difference ($\mathrm{MAD} = \max\{|X_i(n) - \hat{X}_i(n)|: 1 \le i \le N \text{ and } 1 \le n \le L\}$), a third one is the maximum spectral angle

$$\mathrm{MSA} = \max\left\{\operatorname{acos}\left(\frac{\sum_{i=1}^{N}X_i(n)\hat{X}_i(n)}{\sqrt{\sum_{i=1}^{N}X_i^2(n)\sum_{i=1}^{N}\hat{X}_i^2(n)}}\right) : 1 \le n \le L\right\}$$

and the last one is the mean absolute error

$$\left(\mathrm{MAE} = \frac{1}{NL}\sum_{i=1}^{N}\sum_{n=1}^{L}|X_i(n) - \hat{X}_i(n)|\right).$$

With these four distortions, one can estimate the performances of a codec on usual applications of hyperspectral images, like classifications and targets detections [7]. For multispectral images, we considered only the MAD and the MSE distortions, the last one being expressed in terms of PSNR (peak of signal to noise ratio): $\mathrm{PSNR} = 10\log_{10}(2^{N_b} - 1)^2/D$, where N_b is the number of bits per pixel and per band (bpppb) of the initial image. A difference exists between the aimed bit-rate and the actual bit-rate obtained with the VM9. In our tests, this difference does not exceed ± 0.001 bpppb and thus the precision of the PSNR is about ± 0.05 dB.

In our experiments, two kinds of images have been used: hyperspectral and multispectral. The first (*Moffett*, *Cuprite*, *Jasper* shown in Fig. 1) are AVIRIS images,[5] with $N_b = 16$ bpppb, $N = 224$ spectral components from the visible to the infrared spectral light and are originally acquired with $N_r \times N_c = 512 \times 624$, but for the simulations we kept only the 512 leftmost columns. The multispectral images are PLEIADES[6] leftmost columns simulations of French cities with $N = 4$ components, $N_b = 12$ bpppb and their dimensions are: 3152×320 for *Moissac*, 320×1376 for *Port-de-Bouc* (this image has been $\pi/2$ rotated), 3272×352 for *Strasbourg* and 3736×352 for *Vannes*.

On the multispectral images, we observed that JADO performs roughly as the KLT for MSE distortion, sometimes slightly better, sometimes slightly worse, at any rate. On six images, the average gain of JADO on the KLT is negligible (about 0.02 dB) at medium and high bit-rates (from 0.25 to 3 bpppb), whereas the average gain of OrthOST on JADO is about 0.21 dB at the same rates. Nevertheless, on hyperspectral images, JADO performs slightly but significantly better than the KLT for the four distortions tested at medium and high bit-rates (see Table 1) and nearly reaches the OrthOST scores with a significantly lower computational complexity. The average gain of JADO on the KLT (resp. OrthOST on JADO) is 0.37 dB (resp. 0.07 dB) on [0.25 bpppb, 3 bpppb].

[4] *JPEG2000 Verification Model* 9.1 (*Technical description*), ISO/IEC JTC 1/SC 29/WG 1 WG1 N2165, June 2001.

[5] These AVIRIS images have been downloaded from the NASA web site http://aviris.jpl.nasa.gov/html/aviris.freedata.html.

[6] These PLEIADES images have been given by the French Space Agency CNES (Centre National d'Etudes Spatiales), a description of PLEIADES can be found at http://smsc.cnes.fr/PLEIADES/.

Table 1
Bit-rate (in bpppb) versus SNR or PSNR (in dB), versus MAE, versus MAD and versus MSA (in degree °) of different spectral transforms on hyper- and multi-spectral images.

	SNR (dB)								MAE							
Bit-rate	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
	Moffett															
KLT	44.21	47.68	50.08	51.97	54.76	57.10	59.21	61.04	5.36	3.83	3.03	2.49	1.82	1.39	1.07	0.85
JADO	45.13	48.39	50.70	52.50	55.17	57.47	59.53	61.30	4.87	3.54	2.83	2.35	1.74	1.33	1.03	0.82
OrthOST	45.31	48.57	50.87	52.61	55.28	57.57	59.62	61.37	4.77	3.47	2.78	2.32	1.72	1.31	1.02	0.81
	Cuprite															
KLT	47.79	50.46	52.55	54.16	56.76	59.07	61.26	63.27	5.11	3.96	3.23	2.73	2.04	1.55	1.19	0.92
JADO	48.22	50.85	52.86	54.42	56.97	59.27	61.44	63.43	4.86	3.80	3.13	2.65	1.99	1.51	1.16	0.90
OrthOST	48.25	50.88	52.89	54.44	56.99	59.29	61.46	63.44	4.83	3.79	3.12	2.65	1.98	1.51	1.16	0.90
	Jasper															
KLT	42.93	46.49	48.61	50.37	53.18	55.56	57.72	59.66	5.78	4.04	3.27	2.72	1.99	1.51	1.16	0.91
JADO	43.56	46.89	48.97	50.67	53.43	55.78	57.91	59.83	5.42	3.87	3.15	2.63	1.94	1.47	1.13	0.89
OrthOST	43.66	46.94	49.02	50.73	53.47	55.81	57.94	59.85	5.35	3.85	3.13	2.62	1.93	1.46	1.13	0.88
	MSA (°)								MAD							
	Moffett															
KLT	1.43	0.87	0.57	0.37	0.20	0.15	0.12	0.10	392	211	119	67	24	14	8	7
JADO	1.15	0.59	0.42	0.27	0.19	0.14	0.11	0.09	279	120	67	44	18	12	8	6
OrthOST	0.96	0.47	0.31	0.25	0.18	0.14	0.11	0.09	261	77	49	33	18	10	8	6
	Cuprite															
KLT	0.42	0.25	0.22	0.15	0.12	0.08	0.07	0.06	154	135	100	54	26	16	10	8
JADO	0.33	0.25	0.16	0.14	0.10	0.08	0.07	0.05	112	109	61	39	20	12	9	7
OrthOST	0.32	0.25	0.17	0.14	0.10	0.08	0.07	0.05	113	110	61	37	22	11	9	7
	Jasper															
KLT	0.91	0.53	0.43	0.34	0.26	0.20	0.15	0.12	225	151	82	57	30	15	10	7
JADO	0.87	0.51	0.44	0.33	0.24	0.19	0.15	0.12	157	91	56	51	20	11	9	7
OrthOST	0.83	0.51	0.40	0.33	0.24	0.19	0.15	0.12	157	84	46	34	23	13	9	7
	PSNR (dB)								MAD							
	Vannes															
KLT	41.36	45.71	48.78	51.11	54.39	56.82	59.24	61.79	482	219	148	86	51	33	24	18
JADO	41.83	46.15	49.16	51.42	54.61	57.09	59.53	62.06	368	214	134	84	48	29	24	19
OrthOST	41.90	46.27	49.29	51.54	54.70	57.18	59.62	62.16	354	187	135	91	46	30	25	18

5. Conclusion

We studied a compression scheme for multicomponent images that is compatible with the JPEG2000 Part 2 standard and which consists in applying a 2-D DWT on each component and a spectral linear transform between components. In [6], this compression scheme was studied when the spectral transform adapts to the image, while the 2-D DWT was fixed and the criterion (1) satisfied by an optimal transform has been presented. In this paper, we have derived a simplified criterion by treating the transformed coefficients in each subband as having a Gaussian distribution of variance depending on the subband. Then we showed that the transform minimizing the simplified criterion is returned by the algorithm JADO which realizes joint approximate diagonalization of positive definite matrices under orthogonality constraint. Finally, we have compared the performances in coding of this transform with the KLT and the optimal orthogonal transforms obtained without the Gaussianity assumption, for medium to high bit-rates on two kinds of multicomponent images: multispectral and hyperspectral. We have observed that, for hyperspectral images, the transform returned by JADO performs appreciably better than the KLT at medium to high bit-rates. However, this phenomenon is not observed on multispectral images, where the transform returned by JADO performs roughly like the KLT.

Appendix A. The JADO (joint approximate diagonalization under orthogonality constraint) algorithm

Given K positive definite (complex) matrices $\mathbf{C}_1,\dots,\mathbf{C}_K$ associated with positive weights $w_1,\dots,w_K$, the JADO algorithm aims to find a unitary matrix $\mathbf{B}$ which minimizes

$$C(\mathbf{B}) = \sum_{k=1}^{K} w_k \operatorname{logdet} \operatorname{diag}(\mathbf{B}\mathbf{C}_k\mathbf{B}^*), \tag{5}$$

where * denotes the hermitian operator. This algorithm differs only slightly from FG algorithm in [8]. However, its derivation in [8] is complex and difficult to understand.

Here we provide briefly a much simpler derivation. The idea is to make successive Givens rotations, each time on a pair of rows of $\mathbf{B}$, the i th row $\mathbf{B}_{i\cdot}$ and the j th row $\mathbf{B}_{j\cdot}$, say

$$\begin{bmatrix}\mathbf{B}_{i\cdot}\\ \mathbf{B}_{j\cdot}\end{bmatrix} \leftarrow \mathbf{T}_{ij}\begin{bmatrix}\mathbf{B}_{i\cdot}\\ \mathbf{B}_{j\cdot}\end{bmatrix}, \tag{6}$$

where $\mathbf{T}_{ij}$ is a 2×2 unitary matrix, chosen so that the criterion is decreased. The processing of all the $K(K-1)/2$ pairs is called a sweep. The algorithm consists of repeated sweeps until convergence is achieved.

The decrease of the criterion (5) induced by (6) is

$$\sum_{k=1}^{K} w_k \log\left[(\mathbf{B}_{i\cdot}\mathbf{C}_k\mathbf{B}_{i\cdot}^*)(\mathbf{B}_{j\cdot}\mathbf{C}_k\mathbf{B}_{j\cdot}^*)\det \operatorname{diag}\left(\mathbf{T}_{ij}\begin{bmatrix}\mathbf{B}_{i\cdot}\\ \mathbf{B}_{j\cdot}\end{bmatrix}\mathbf{C}_k[\mathbf{B}_{i\cdot}^* \quad \mathbf{B}_{j\cdot}^*]\mathbf{T}_{ij}^*\right)\right].$$

A natural idea is to chose $\mathbf{T}_{ij}$ to maximize this decrease, but there is no closed form formulae for that. Our idea is to maximize a lower bound of it instead. Since for $a>0$, $b\geq 0$, $\log(a/b)\geq 1-b/a$, the above decrease can be seen to be bounded below by

$$2(w_1+\cdots+w_K)-\mathbf{T}_{ij;1\cdot}\mathbf{P}\mathbf{T}_{ij;1\cdot}^*-\mathbf{T}_{ij;2\cdot}\mathbf{Q}\mathbf{T}_{ij;2\cdot}^*, \tag{7}$$

where $\mathbf{T}_{ij;1\cdot}$ and $\mathbf{T}_{ij;2\cdot}$ are the first and second rows of $\mathbf{T}_{ij}$ and

$$\mathbf{P}=\sum_{k=1}^{K}\frac{w_k}{\mathbf{B}_{i\cdot}\mathbf{C}_k\mathbf{B}_{i\cdot}^*}\begin{bmatrix}\mathbf{B}_{i\cdot}\\ \mathbf{B}_{j\cdot}\end{bmatrix}\mathbf{C}_k[\mathbf{B}_{i\cdot}^* \quad \mathbf{B}_{j\cdot}^*],$$

$$\mathbf{Q}=\sum_{k=1}^{K}\frac{w_k}{\mathbf{B}_{j\cdot}\mathbf{C}_k\mathbf{B}_{j\cdot}^*}\begin{bmatrix}\mathbf{B}_{i\cdot}\\ \mathbf{B}_{j\cdot}\end{bmatrix}\mathbf{C}_k[\mathbf{B}_{i\cdot}^* \quad \mathbf{B}_{j\cdot}^*].$$

Since $\mathbf{T}_{ij;2\cdot}$ has unit norm and is orthogonal to $\mathbf{T}_{ij;1\cdot}$, it must be of the form $e^{i\alpha}\overline{\mathbf{T}}_{ij;1\cdot}\mathbf{J}$ where α is some phase angle, $\overline{x}$ denotes the complex conjugate of x and $\mathbf{J}$ is the 2×2 matrix with 0 on the diagonal and 1, -1 on the anti-diagonal. Thus $\mathbf{T}_{ij;2\cdot}\mathbf{Q}\mathbf{T}_{ij;2\cdot}^*=\overline{\mathbf{T}}_{ij;1\cdot}\mathbf{J}\mathbf{Q}\mathbf{J}^*\overline{\mathbf{T}}_{ij;1\cdot}^*$, but since the above left hand side is real (as $\mathbf{Q}$ is hermitian), it also equals $\mathbf{T}_{ij;1\cdot}\mathbf{J}\overline{\mathbf{Q}}\mathbf{J}^*\mathbf{T}_{ij;1\cdot}^*$. Therefore expression (7) can be rewritten as $2w-\mathbf{T}_{ij;1\cdot}(\mathbf{P}+\mathbf{J}\overline{\mathbf{Q}}\mathbf{J}^*)\mathbf{T}_{ij;1\cdot}^*$. Maximizing it with respect to the unitary matrix $\mathbf{T}$ thus amounts to minimizing $\mathbf{T}_{ij;1\cdot}(\mathbf{P}+\mathbf{J}\overline{\mathbf{Q}}\mathbf{J}^*)\mathbf{T}_{ij;1\cdot}^*$ with respect to the vector of unit norm $\mathbf{T}_{ij;1\cdot}$. The solution is that $\mathbf{T}_{ij;1\cdot}$ is (up to a factor of unit modulus) the normalized left eigenvector of the smallest eigenvalue of $\mathbf{P}+\mathbf{J}\overline{\mathbf{Q}}\mathbf{J}^*$. Since $\mathbf{T}_{ij;2\cdot}$ is orthogonal to $\mathbf{T}_{ij;1\cdot}$ it is the other eigenvector. Finally, $\mathbf{T}_{ij}$ is the matrix formed by the left eigenvectors of $\mathbf{P}+\mathbf{J}\overline{\mathbf{Q}}\mathbf{J}^*$. Its elements can be computed explicitly in closed form.

We note that the off diagonal elements of $\mathbf{J}\overline{\mathbf{Q}}\mathbf{J}^*$ is the negative of those of $\mathbf{Q}$ while the diagonal elements are those of $\mathbf{Q}$ *in reverse order*. Thus $\mathbf{J}\overline{\mathbf{Q}}\mathbf{J}^*=\operatorname{tr}(\mathbf{Q})\mathbf{I}-\mathbf{Q}$ where tr denotes the trace. Since the addition of a multiple of the identity matrix does not change the eigenvectors, $\mathbf{T}_{ij}$ is also the matrix formed by the left eigenvectors of $\mathbf{P}-\mathbf{Q}$. One can now recognize that the rotation (6) is the same as an iteration in the G loop of the FG algorithm. However, it differs from our JADO algorithm in that it repeats (6) with the same pair i,j (but with the newly computed $\mathbf{B}_{i\cdot}$ and $\mathbf{B}_{j\cdot}$) until convergence (the G loop) and only then another pair i,j is considered. We feel that this is not efficient since the decrease of the criterion will be very small near the end of the G loop.

References

[1] D.S. Taubman, M.W. Marcellin, JPEG2000: Image Compression Fundamentals, Standards and Practice, Kluwer Academic Publishers, Dordrecht, 2002.

[2] J.E. Fowler, J.T. Rucker, 3D wavelet-based compression of hyperspectral imagery, in: C.-I. Chang (Ed.), Hyperspectral Data Exploitation: Theory and Applications, Wiley, Hoboken, NJ, 2007 (Chapter 14).

[3] E. Christophe, C. Mailhes, P. Duhamel, Best anisotropic 3-D wavelet decomposition in a rate-distortion sense, in: Proceedings of the IEEE ICASSP'06, vol. 2, May 2006, II-17–20.

[4] Q. Du, J.E. Fowler, Hyperspectral image compression using JPEG2000 and principal component analysis, IEEE Geoscience and Remote Sensing Letters 4 (April 2007) 201–205.

[5] B. Penna, T. Tillo, E. Magli, G. Olmo, Transform coding techniques for lossy hyperspectral data compression, IEEE Transactions on Geoscience and Remote Sensing 45 (5) (May 2007).

[6] I.P. Akam Bita, M. Barret, D.-T. Pham, On optimal transforms in lossy compression of multicomponent images with JPEG2000, Signal Processing, accepted for publication, doi:10.1016/j.sigpro.2009.09.011.

[7] E. Christophe, D. Léger, C. Mailhes, Quality criteria benchmark for hyperspectral imagery, IEEE Transactions on Geoscience and Remote Sensing 43 (9) (September 2005) 2103–2114.

[8] B.N. Flury, W. Gautschi, An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form, SIAM Journal on Scientific and Statistical Computing 7 (1) (January 1986).

[9] A. Gersho, R.M. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publisher, 1992.

[10] A. Papoulis, Probability, Random Variables, and Stochastic Processes, 2nd ed., McGraw- Hill, 1984.

Signal Processing 90 (2010) 759–773

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Review

On optimal transforms in lossy compression of multicomponent images with JPEG2000 ☆

Isidore Paul Akam Bita [a,*], Michel Barret [b,1], Dinh-Tuan Pham [c,2]

[a] LUXSPACE Sarl, Chateau de Betzdorf, L-6815 Betzdorf, Luxembourg
[b] SUPELEC, Information Multimodality and Signal Team, 2 rue É. Belin 57070 Metz, France
[c] Jean Kuntzmann Laboratory, 51 rue des Mathématiques, BP 53, 38041 Grenoble Cedex 9, France

ARTICLE INFO

Article history:
Received 15 October 2008
Received in revised form
22 May 2009
Accepted 11 September 2009
Available online 22 September 2009

Keywords:
Multicomponent image coding
Multicomponent image compression
High rate transform coding
Karhunen–Loève transform
Independent component analysis
JPEG2000
Multispectral image codind
Hyperspectral image coding
Optimal transform coding

ABSTRACT

It is well known in transform coding, that the Karhunen–Loève transform (KLT) is optimal only for Gaussian sources. However, in many applications using JPEG2000 Part 2 codecs, the KLT is generally considered as the optimal linear transform for reducing redundancies between components of multicomponent images. In this paper we present the criterion satisfied by an optimal transform of a JPEG2000 compatible compression scheme, under high resolution quantization hypothesis and without the Gaussianity assumption. We also introduce two variants of the compression scheme and the associated criteria minimized by optimal transforms. Then we give two algorithms, derived of the Independent Component Analysis algorithm `ICAinf`, that compute the optimal transform, one under the orthogonality constraint and the other without no constraint but invertibility. The computational complexity of the algorithms is evaluated. Finally, comparisons with the KLT are presented on hyperspectral and multispectral satellite images with different measures of distortion, as it is recommended for evaluating the performances of the codec in applications (like classification and target detection). For hyperspectral images, we observe a little but significant gain at medium and high bit-rates of the optimal transforms compared to the KLT. The actual drawback of the optimal transforms is their heavy computational complexity.

Contents

☆ This work was supported in part by the French Ministry of Higher Education and Research, with the "Action Concertée Incitative Masse de Données" ACI^{2}M project.
* Corresponding author. Tel.: +352 267 890 8046; fax: +352 267 890 4029.
E-mail addresses: AkamBita@luxspace.lu (I.P. Akam Bita), Michel.Barret@supelec.fr (M. Barret), Dinh-Tuan.Pham@imag.fr (D.-T. Pham).
[1] Tel.: +33 38 77 64731; fax: +33 38 77 64700.
[2] Tel.: +33 47 65 14423; fax: +33 47 66 31263.

0165-1684/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2009.09.011

1. Introduction

These last years, research activities on multicomponent image compression have been expanded, due to the development of multispectral and hyperspectral image sensors which supply larger and larger amount of data. The end-users of such images are also more and more numerous and miscellaneous. The future earth observation systems, for instance, will use multi-, hyper- (even super-) spectral image sensors with higher resolutions leading to bigger amount of transmitted data. However, the channel bandwidth for transmission is limited and therefore it is of great interest to conceive compression systems (onboard and on the ground) of multicomponent images compatible with the diversity of end-users' needs.

The components of a multicomponent image generally represent the same scene with different views depending on the wavelength. Therefore there is a high degree of dependence (or redundancies) between components, and a particularity of multi- or hyperspectral images is that there exist two kinds of redundancies: spatial (between the different pixels in each component) and spectral (between the components).

During the past decades, different solutions have been proposed for multicomponent image compression. A solution currently adopted consists of using two different transformations, each with the goal of reducing only one of the two redundancies. In [1], a 2-D discrete wavelet transform (DWT) is used to reduce the spatial redundancies in each component while the Karhunen Loève transform (KLT) is applied to reduce the spectral ones. In that paper, the quantization and entropy coding are achieved thanks to the well known SPIHT (set partitioning in hierarchical trees) codec by Said and Pearlman [2] in its original version and in a modified version including VQ (vector quantization). In the same way, with the use of the 2-D DWT of [4] (usually called the Daubechies 9/7), the authors of [3] use a lattice VQ with a stack run coder as quantization and entropy coding. More recently in [5], the KLT associated with the Daubechies 9/7 2-D DWT and with EBCOT [6,7] for quantizing and entropy coding has been tested on hyperspectral images with different bit-allocations between components. It is shown that the post-compression rate-distortion (PCRD) optimizer of EBCOT applied across multiple bands gives the best rate-distortion performance. Another solution consists of using a 3-D DWT for reducing both the spatial and spectral redundancies with only one transform. This approach is generally applied to hyperspectral images as in [8]. An overview of 3-D wavelet-based techniques and more can be found in [10]. The two abovementioned solutions are compatible with the JPEG2000 Part 2 standard.

The JPEG2000 standard is well known and well spread today. Moreover the KLT used in JPEG2000 Part 2 is considered as the best existing lossy compression techniques for hyperspectral images at medium and high bit rates [11,12]. The KLT consists of a principal component analysis (PCA), well known of statisticians, where all the components are kept. However, the rather great computational complexity of the KLT hinders its adoption in practice—specially on satellite platforms—and recent works propose different solutions in order to pass round this problem. One approach consists in reducing the complexity of the covariance matrix computation. This is done by randomly sampling the entire image in order to obtain a small sample of the pixels' population on which the covariance matrix is computed [12,13]. Another approach consists in computing a kind of KLT average on a set of images (the learning basis) issued from only one sensor and using it on other images obtained with the same sensor. This sub-optimal transform is called exogenous KLT in [14] and the computational complexity of the second approach is compatible with satellite platforms. Both approaches are fruitful: the rate-distortion performance sacrifice compared with the true KLT is very slight, whereas the computational burden is significantly reduced. In the second approach, the exogenous KLT matrix is known by the decoder, hence there is no need to transmit it.

It is often taught in coding lessons that the KLT is optimal in transform coding. Nevertheless it is well known that the optimality of the KLT is proven only for Gaussian data [15–17] and that it can be sub-optimal for non-Gaussian data [18]. Now, under the high resolution quantization hypothesis, nearly everything is known

about the performance of a transform coding with entropy constrained scalar quantization and mean square distortion. It is then straightforward to find a criterion that, when minimized, gives the optimal linear transform under the above mentioned conditions. Nevertheless, the optimal transform computation is generally considered as a difficult task [19] and the Gaussian assumption is then used in order to simplify the calculus. In [20] the problem of computing the optimal transform for still images is resolved under high-rate entropy constraint scalar quantization hypothesis.

In this paper, we clarify the criterion minimized by an optimal linear transform for reducing spectral redundancies under high-rate entropy constraint scalar quantization hypothesis and mean square error distortion, when a 2-D DWT—with fixed coefficients—is applied to each component to reduce spatial redundancies and one scalar quantizer per subband and per component is used. This compression scheme is compatible with the JPEG2000 Part 2 standard. We extend it to two variants which are not JPEG2000 compatible. We show the link between the criterion and the mutual information contrast used in independent component analysis (ICA). Moreover, for each case we describe the quasi-Newton algorithms used for the minimization of the criterion, either with the constraint of an orthogonal transform or without no constraint but invertibility. These algorithms are derived from an algorithm by Pham `ICAinf` described in [24] that performs ICA. Finally, performances of these transforms and comparisons with the KLT are given for multi- and hyperspectral satellite images, with different measures of distortion: signal to noise ratio (SNR), maximum absolute difference (MAD), mean absolute error (MAE) and maximum spectral angle (MSA). Indeed, it is well-known that providing the mean square error as the only estimate of distortion is not sufficient to assess the quality of a codec for hyperspectral images [25]. It is important to notice that the method abovementioned, which consists in computing an exogenous KLT on a learning basis of images, has been successfully applied to the new optimal transforms described in the next sections, leading to a codec with an acceptable complexity for satellite platforms [9].

In the next section we present the compression scheme with two improved variants. Then in Section 3 we recall the asymptotical expression of the mean square error distortion under high-rate entropy constraint scalar quantization hypothesis and consequently we clarify in Section 4 the criterion minimized by an optimal linear transform under the same assumptions. In Section 5 we describe the quasi-Newton algorithms that return the optimal transforms and finally performances in coding and comparison with the KLT are shown and analyzed in the last section. The content of this paper has been partially published in [21,22].

2. Description of the compression scheme with two variants

2.1. Conventions and notations

We consider a multicomponent image $\mathbf{X}$ with N components $\mathbf{X}_1,\dots,\mathbf{X}_N$. Each component $\mathbf{X}_i$ is a 2-D image with N_r rows and N_c columns. To simplify the notations and the mathematical expressions, we assume that each component is written as a row vector by scanning all its pixels row by row (for example). Then $\mathbf{X}$ is an $N\times L$ matrix, with $L=N_rN_c$. In the following, depending on the context, we shall interpret $\mathbf{X}_i$ as a 2-D image or as a row vector of dimension L.

We write vec to denote the linear operator that transforms a matrix into a column vector by stacking up its columns one after the other (e.g., if $\mathbf{a}_1,\dots,\mathbf{a}_K$ are the column vectors of a matrix $\mathbf{A}$, then vec($\mathbf{A}$) is the vector $(\mathbf{a}_1^T,\dots,\mathbf{a}_K^T)^T$, where ^{T} denotes transposition). Let P be an integer and KP a multiple of P. We write perm to denote the linear operator that transforms an $N\times KP$ matrix $\mathbf{B}=[\mathbf{B}_1,\mathbf{B}_2,\dots,\mathbf{B}_K]$ (where each block $\mathbf{B}_i$ is of dimension $N\times P$) into the matrix $[\mathrm{vec}(\mathbf{B}_1),\dots,\mathrm{vec}(\mathbf{B}_K)]$ of dimension $NP\times K$. So, if the vector $\mathbf{b}_{i,j}$ denotes the j th column of $\mathbf{B}_i$ ($1\le i\le K$ and $1\le j\le P$), then the i th column of perm($\mathbf{B}$) is the vector $(\mathbf{b}_{i,1}^T,\dots,\mathbf{b}_{i,P}^T)^T$. The mathematical inverse of perm is written as perm^{-1}. As a direct consequence of these definitions, for P vectors $\mathbf{x}_u\in\mathbb{R}^N$ ($1\le u\le P$), the vector $\mathbf{x}=(\mathbf{x}_1^T\dots\mathbf{x}_P^T)^T\in\mathbb{R}^{NP}$ satisfies

$$\mathrm{perm}^{-1}(\mathbf{x}\mathbf{e}_k'^T)=\sum_{u=1}^{P}\mathbf{x}_u\mathbf{e}_{(k-1)P+u}^T, \tag{1}$$

where $\mathbf{e}_v$, for $1\le v\le KP$, and $\mathbf{e}_k'$, for $1\le k\le K$, are, respectively, the v th and the k th vectors of the canonical basis of $\mathbb{R}^{KP}$ and $\mathbb{R}^K$. For a square matrix $\mathbf{M}$, the expressions det$\mathbf{M}$ and diag($\mathbf{M}$) denote, respectively, its determinant and the diagonal matrix obtained with its diagonal elements.

In the following compression schemes, the 2-D DWT have all fixed coefficients (in our tests, the Daubechies 9/7 DWT is always used), but the spectral linear transforms are adapted to the data. We denote $\mathbf{W}$ the invertible $L\times L$ matrix associated with the 2-D DWT.

2.2. The separable scheme

The separable scheme is compatible with the JPEG2000 Part 2 standard. It can be described as follows:

- *Coding*. The same 2-D DWT is applied to each component $\mathbf{X}_i$ in order to reduce the spatial redundancies and a linear transform $\mathbf{A}$ is applied between the components in order to reduce the spectral redundancies. The result of the 2-D DWT applied to the entire image $\mathbf{X}$ is $\mathbf{X}\mathbf{W}^T$ and the transformed coefficients are the elements of the matrix $\mathbf{Y}=\mathbf{A}\mathbf{X}\mathbf{W}^T$. Then, the transformed coefficients are quantized (see Section 6.1) and entropy coded.
- *Decoding*. Let $\mathbf{Y}^q$ denote the matrix with the same dimension as $\mathbf{Y}$ containing the dequantized transformed coefficients. The mathematical inverse transforms are applied to $\mathbf{Y}^q$ in order to reconstruct an approximation $\hat{\mathbf{X}}=\mathbf{A}^{-1}\mathbf{Y}^q\mathbf{W}^{-T}$ of the original image $\mathbf{X}$.

We can remark that the order of the transformations (i.e., applying first the DWT then $\mathbf{A}$, or first $\mathbf{A}$ then the DWT)

has no effect on the result, since $\mathbf{Y} = \mathbf{A}(\mathbf{X}\mathbf{W}^T) = (\mathbf{A}\mathbf{X})\mathbf{W}^T$. This is why that scheme is called *separable*.

2.3. Two non-separable variants of the separable scheme

It is well known that wavelet coefficient statistics depend on the subband. In the separable scheme, if the spectral transform adapts to the data, it cannot fully take into account this difference of statistics, since the same transform is applied to all the subbands. The following scheme rectifies this shortcoming.

2.3.1. The subband scheme

The subband scheme is not compatible with the JPEG2000 Part 2 standard. For that scheme, the DWT is first applied to each component, then spectral linear transforms are applied, with a specific one for each subband. This explains the name "subband scheme". It is clear that the above order of the transformations (DWT first then spectral transforms) must be respected. This scheme can be summarized as follows:

- *Coding*. First, the 2-D DWT is applied to each component $\mathbf{X}_i$ and the wavelet coefficients are the elements of the matrix $\mathbf{X}\mathbf{W}^T$. Second, for each component, the wavelet coefficients of each subband are regrouped according to a fixed scan that does not depend on the component. This re-ordering corresponds to the right multiplication of $\mathbf{X}\mathbf{W}^T$ by a permutation matrix $\mathbf{P}^T$. We can suppose without loss of generality that $\mathbf{P}$ is the identity, otherwise we could replace $\mathbf{W}$ with $\mathbf{P}\mathbf{W}$. This partitioning can be written as $\mathbf{X}\mathbf{W}^T = [(\mathbf{X}\mathbf{W}^T)^{(1)} \dots (\mathbf{X}\mathbf{W}^T)^{(M)}]$, where M is the number of subbands per component. Then M spectral transforms are applied, one per subband ($\mathbf{A}^{(m)}$ for subband m), and the transformed coefficients are the elements of the matrix: $\mathbf{Y} = [\mathbf{A}^{(1)}(\mathbf{X}\mathbf{W}^T)^{(1)} \dots \mathbf{A}^{(M)}(\mathbf{X}\mathbf{W}^T)^{(M)}]$. Finally the transformed coefficients are quantized and entropy coded.
- *Decoding*. Let $\mathbf{Y}^q = [\mathbf{Y}^{q(1)}, \dots, \mathbf{Y}^{q(M)}]$ denote the matrix of the dequantized transformed coefficients, regrouped by subbands. The mathematical inverse transforms are applied to $\mathbf{Y}^q$ in order to reconstruct an approximation $\hat{\mathbf{X}} = [\mathbf{A}^{(1)^{-1}}\mathbf{Y}^{q(1)}, \dots, \mathbf{A}^{(M)^{-1}}\mathbf{Y}^{q(M)}]\mathbf{W}^{-T}$ of the original image $\mathbf{X}$.

2.3.2. The mixed subband scheme

It is well known that after a DWT some redundancies remain between adjoining wavelet coefficients [23]. The entropy coder EBCOT of JPEG2000 exploits these redundancies [6]. In the subband scheme, the spectral transforms reduce only the spectral redundancies and leave the redundancies between wavelet coefficients unchanged. The mixed subband scheme attempts to rectify this shortcoming. The difference with the subband scheme consists of using spectral linear transforms in order to reduce at once the redundancies between components (i.e., spectral ones) and the redundancies between adjoining wavelet coefficients localized in a same subband. This is done by adding a stage of polyphase decomposition with P-fold decimators (realized with the operator perm) on wavelet coefficients (see Fig. 1). The name of this scheme originates from mixing of redundancies reduction (spectral and spatial) with "spectral" transforms (one per subband). In order to apply the operator perm to each subband of each component, we suppose that the cardinal number of each subband is divisible by P and that $L = KP$. The mixed subband scheme can be summarized as follows:

- *Coding*. First, the DWT is applied to each component $\mathbf{X}_i$ and the wavelet coefficients are the elements of the matrix $\mathbf{X}\mathbf{W}^T$. Second, for each component, the wavelet coefficients of each subband are regrouped according to a fixed scan that does not depend on the component—as in the subband scheme. This partitioning can be written as $\mathbf{X}\mathbf{W}^T = [(\mathbf{X}\mathbf{W}^T)^{(1)} \dots (\mathbf{X}\mathbf{W}^T)^{(M)}]$, where M is the number of subbands in a component. The abovementioned scan depends on P: it must transform P adjoining wavelet coefficients into P successive elements. Third, the operator perm is applied to each subband:

$$[(\mathbf{X}\mathbf{W}^T)^{(1)} \dots (\mathbf{X}\mathbf{W}^T)^{(M)}] \mapsto [\mathrm{perm}\{(\mathbf{X}\mathbf{W}^T)^{(1)}\} \dots \mathrm{perm}\{(\mathbf{X}\mathbf{W}^T)^{(M)}\}].$$

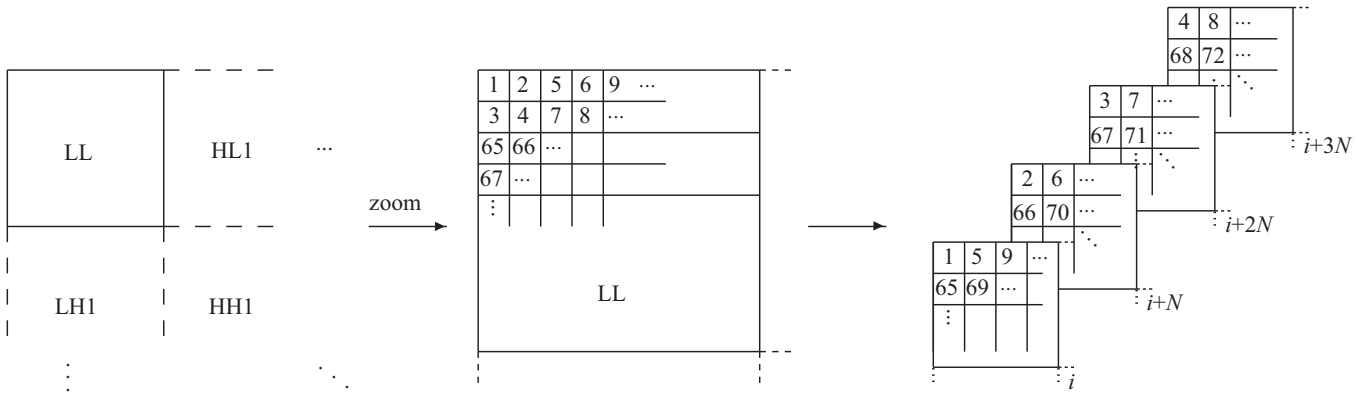


Fig. 1. Mixed subband scheme: each component is splitted into P components (the i th-component gives the components $i, i+N, \dots, i+(P-1)N$), here $P = 4$ and the LL subband is of dimension 32×32.

Then, spectral linear transforms depending on the subband are applied, one per subband. The transform $\mathbf{A}^{(m)}$ associated with the subband m is an invertible $NP \times NP$ matrix and the transformed coefficients are the elements of the matrix $\mathbf{Y} = [\mathbf{A}^{(1)}\text{perm}\{(\mathbf{X}\mathbf{W}^T)^{(1)}\} \dots \mathbf{A}^{(M)}\text{perm}\{(\mathbf{X}\mathbf{W}^T)^{(M)}\}]$ of dimension $NP \times K$. Finally the transform coefficients are quantized and entropy coded.

- *Decoding*. From $[\mathbf{Y}^{q(1)} \dots \mathbf{Y}^{q(M)}]$, the dequantized transformed coefficients regrouped subband by subband, the inverse transforms $\mathbf{A}^{(m)^{-1}}$ are applied for $1 \le m \le M$ giving $[\mathbf{A}^{(1)^{-1}}\mathbf{Y}^{q(1)} \dots \mathbf{A}^{(M)^{-1}}\mathbf{Y}^{q(M)}]$. Then the N components are reconstructed with perm^{-1}:

$$[\mathbf{A}^{(1)^{-1}}\mathbf{Y}^{q(1)} \dots \mathbf{A}^{(M)^{-1}}\mathbf{Y}^{q(M)}] \mapsto [\text{perm}^{-1}\{\mathbf{A}^{(1)^{-1}}\mathbf{Y}^{q(1)}\} \dots \text{perm}^{-1}\{\mathbf{A}^{(M)^{-1}}\mathbf{Y}^{q(M)}\}].$$

Finally the inverse DWT is applied to each component, giving the approximation $\hat{\mathbf{X}}$:

$$\hat{\mathbf{X}} = [\text{perm}^{-1}\{\mathbf{A}^{(1)^{-1}}\mathbf{Y}^{q(1)}\} \dots \text{perm}^{-1}\{\mathbf{A}^{(M)^{-1}}\mathbf{Y}^{q(M)}\}]\mathbf{W}^{-T}.$$

We denote K_m the number of columns of $\mathbf{Y}$ associated with the subband m.

We can remark that the subband scheme is a special case of the mixed subband scheme obtained when $P = 1$ and that the separable scheme is a special case of the subband scheme obtained when all the spectral transforms $\mathbf{A}^{(m)}$ are the same.

3. Expressions of the distortion

The problem of bit allocation between quantizers is well solved when the end-to-end distortion between $\mathbf{X}$ and $\hat{\mathbf{X}}$ is well approximated by an (eventually weighted) sum of the quantizers' distortions [26]. Therefore it is important to have such a good approximation. When the entire 3-D transformation consists in associations of 2-D DWT and spectral linear transforms, as in the compression schemes of the previous section, these approximations are not directly available in the literature (to our knowledge); however, they are straightforward to deduce from well-known results [26,17,6]. In this section we give the relation that links the distortion between $\mathbf{X}$ and $\hat{\mathbf{X}}$ to the quantizers' distortions, when the distortion is the mean square error. We begin by recalling the solution of the problem in a simple general case [26,17,6].

3.1. A simple general case

Lemma 3.1. *Let $\mathbf{X}$ be a real random vector with N components and $\mathcal{A}$ be an invertible matrix of order N. The transformed vector $\mathbf{Y} = \mathcal{A}\mathbf{X}$ is quantized and dequantized in $\mathbf{Y}^q$. The original vector $\mathbf{X}$ is approximated by $\hat{\mathbf{X}} = \mathcal{A}^{-1}\mathbf{Y}^q$ and let $\mathbf{b} = \mathbf{Y} - \mathbf{Y}^q$ be the quantization noise. Then, the end-to-end distortion $D = (1/N)E(\|\mathbf{X} - \hat{\mathbf{X}}\|^2)$, where E denotes the mathematical expectation, satisfies the relation $D = (1/N)\,\text{tr}[E(\mathbf{b}\mathbf{b}^T)\mathcal{A}^{-T}\mathcal{A}^{-1}]$, where* tr *is the trace operator.*

Proof. We have $\mathbf{X} - \hat{\mathbf{X}} = \mathcal{A}^{-1}\mathbf{b}$ and $\|\mathcal{A}^{-1}\mathbf{b}\|^2 = \mathbf{b}^T\mathcal{A}^{-T}\mathcal{A}^{-1}\mathbf{b} = \text{tr}[\mathcal{A}^{-1}\mathbf{b}\mathbf{b}^T\mathcal{A}^{-T}] = \text{tr}[\mathbf{b}\mathbf{b}^T\mathcal{A}^{-T}\mathcal{A}^{-1}]$, therefore $D = (1/N)E(\|\mathcal{A}^{-1}\mathbf{b}\|^2) = (1/N)\text{tr}[E(\mathbf{b}\mathbf{b}^T)\mathcal{A}^{-T}\mathcal{A}^{-1}]$. □

Further, we may need the following assumption, that can be deduced from high resolution quantization hypothesis [26] (this point is recalled in Appendix A).

$\mathcal{H}_1$: *the components of the quantization noise are zero mean and uncorrelated.*

Theorem 1. 1. *With the notations and hypotheses of Lemma* 3.1 *and assuming $\mathcal{H}_1$, the distortion becomes*

$$D = \frac{1}{N}\sum_{i=1}^{N} w_i D_i, \tag{2}$$

where $D_i = E(b_i^2)$ is the quantizer distortion of the ith component $\mathbf{Y}_i$ of $\mathbf{Y}$ and

$$w_i = \sum_{j=1}^{N} (\mathcal{A}^{-1})_{ji}^2 = \|\mathcal{A}^{-1}\mathbf{e}_i\|^2, \tag{3}$$

with $(\mathcal{A}^{-1})_{ij}$ the element of $\mathcal{A}^{-1}$ located on row i and column j and $\mathbf{e}_i$ the ith canonical vector of $\mathbb{R}^N$.

The relations (2)–(3) *hold without the assumption $\mathcal{H}_1$ if $\mathcal{A}^{-T}\mathcal{A}^{-1}$ is diagonal, e.g., if $\mathcal{A}$ is orthogonal.*

Proof. The assumptions in 1 or 2 state that at least one of the two matrices $\mathcal{A}^{-T}\mathcal{A}^{-1}$ and $E(\mathbf{b}\mathbf{b}^T)$ is diagonal. Hence the trace of their product is equal to the sum of the products of their diagonal elements. □

3.2. The mixed subband scheme

In the following, the symbols $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Y}^q$ refer again to the matrices defined in Section 2 and $\mathcal{A}$ denotes the matrix of the linear transform that associates $\mathbf{Y}$ with $\mathbf{X}$. We are going to apply the formulae of the general simplified case to the mixed subband scheme. In order to express the relation (3) in terms of the DWT $\mathbf{W}$ and the spectral transforms $\mathbf{A}^{(m)}$, it is important to note first that the canonical basis of the space of matrices of dimension $NP \times K$ is the family of matrices $\mathbf{e}_{i,k} = \mathbf{e}_i\mathbf{e}_k'^T$ $(1 \le i \le NP,\ 1 \le k \le K)$, with $\mathbf{e}_i$ (resp. $\mathbf{e}_k'$) the i th (resp. k th) vector of the canonical basis of $\mathbb{R}^{NP}$ (resp. $\mathbb{R}^K$). If k $(1 \le k \le K)$ refers to a column index of the matrix $\mathbf{Y}$ located in the subband m $(1 \le m \le M)$, then using the relation (1) we have

$$\mathcal{A}^{-1}\,\text{vec}(\mathbf{e}_{i,k}) = \text{vec}\{[\text{perm}^{-1}(\mathbf{A}^{(m)^{-1}}\mathbf{e}_i\mathbf{e}_k'^T)]\mathbf{W}^{-T}\} = \text{vec}\left\{\sum_{u=1}^{P}[\mathbf{A}^{(m)^{-1}}\mathbf{e}_i]_u(\mathbf{W}^{-1}\mathbf{e}_{(k-1)P+u})^T\right\}. \tag{4}$$

In this expression, $\mathbf{e}_{(k-1)P+u}$ is the $[(k-1)P+u]$th vector of the canonical basis of $\mathbb{R}^{KP}$ and $\mathbf{A}^{(m)^{-1}}\mathbf{e}_i$ is a vector of size NP that can be partitioned into P non-overlapping blocks of dimension N. We refer to $[\mathbf{A}^{(m)^{-1}}\mathbf{e}_i]_u$ its uth block which is a column vector of size N. Therefore, we can write

$$\|\mathcal{A}^{-1}\,\text{vec}(\mathbf{e}_{i,k})\|^2 = \text{tr}\left\{\left[\sum_{u=1}^{P}[\mathbf{A}^{(m)^{-1}}\mathbf{e}_i]_u(\mathbf{W}^{-1}\mathbf{e}_{(k-1)P+u})^T\right]\right.$$

$$\times\left[\sum_{u=1}^{P}[\mathbf{A}^{(m)^{-1}}\mathbf{e}_i]_u(\mathbf{W}^{-1}\mathbf{e}_{(k-1)P+u})^T\right]^T\Bigg\}$$
$$=\sum_{u,v=1}^{P}(\mathbf{W}^{-1}\mathbf{e}_{(k-1)P+u})^T(\mathbf{W}^{-1}\mathbf{e}_{(k-1)P+v})$$
$$\times\mathrm{tr}\{[\mathbf{A}^{(m)^{-1}}\mathbf{e}_i]_u[\mathbf{A}^{(m)^{-1}}\mathbf{e}_i]_v^T\}$$
$$=\sum_{u,v=1}^{P}\mathbf{e}_{(k-1)P+u}^T\mathbf{W}^{-T}\mathbf{W}^{-1}\mathbf{e}_{(k-1)P+v}[\mathbf{A}^{(m)^{-1}}\mathbf{e}_i]_v^T\times[\mathbf{A}^{(m)^{-1}}\mathbf{e}_i]_u. \quad (5)$$

This last expression is quite complex; however, it can be simplified in some cases: (1) when $P=1$, i.e., for the subband scheme—that will be handled later—and (2) for the mixed subband scheme with the following assumption (which is valid e.g., when the DWT is orthogonal):

$\mathcal{H}_2$: *the matrix* $\mathbf{W}^{-T}\mathbf{W}^{-1}$ *is diagonal.*

We can remark that the condition $\mathcal{H}_2$ is a strong one: only the Haar wavelet is orthogonal, symmetrical and with a compact support. In our tests, we always used the Daubechies 9/7 wavelet, which is almost orthogonal: that is $\mathbf{W}^{-T}\mathbf{W}^{-1}$ is almost diagonal (the non-diagonal elements can be negligible compared to the diagonal ones).

Under assumption $\mathcal{H}_2$, the relation (5) becomes

$$\|\mathcal{A}^{-1}\mathrm{vec}(\mathbf{e}_{i,k})\|^2=\sum_{u=1}^{P}\|\mathbf{W}^{-1}\mathbf{e}_{(k-1)P+u}\|^2\times\|[\mathbf{A}^{(m)^{-1}}\mathbf{e}_i]_u\|^2. \quad (6)$$

Moreover, if we assume that

$\mathcal{H}_3$: *the quantity* $\|\mathbf{W}^{-1}\mathbf{e}_{(k-1)P+u}\|^2$ *does not depend on the spatial position in the subband,*

then the previous equality becomes

$$\|\mathcal{A}^{-1}\mathrm{vec}(\mathbf{e}_{i,k})\|^2=\|\mathbf{W}^{-1}\mathbf{e}_{kP}\|^2\times\|\mathbf{A}^{(m)^{-1}}\mathbf{e}_i\|^2. \quad (7)$$

Remark 1. The condition $\mathcal{H}_3$ is satisfied by dyadic wavelets having finite impulse response (FIR) synthesis filters, when edge effects are neglected (for more details see e.g., [27,28]).

Let us state these results in the following theorem.

Theorem 2. 1. *With the notations of Section* 2.3.2 *and under the assumptions* $\mathcal{H}_1$, $\mathcal{H}_2$ *and* $\mathcal{H}_3$, *the end-to-end distortion verifies*

$$D=\frac{1}{NPK}\sum_{j=1}^{N}\sum_{\ell=1}^{P}\sum_{m=1}^{M}\sum_{k}\|\mathbf{W}^{-1}\mathbf{e}_{kP}\|^2\|\mathbf{A}^{(m)^{-1}}\mathbf{e}_{(\ell-1)N+j}\|^2D_{j,\ell,k}. \quad (8)$$

In the summation, the index k ranges in all the columns of matrix $\mathbf{Y}$ *which are located in the subband m and* $D_{j,\ell,k}$ *is the quantization distortion associated to the* ℓ*th transformed coefficient of the kth block of size P located in the subband m of the jth component.*

2. *When the DWT* $\mathbf{W}$ *and the spectral transforms* $\mathbf{A}^{(m)}$ *are all orthogonal, the relation* (8) *remains valid without the assumption* $\mathcal{H}_1$ *and in this case we have* $\|\mathbf{W}^{-1}\mathbf{e}_{kP}\|^2=\|\mathbf{A}^{(m)^{-1}}\mathbf{e}_{(\ell-1)N+j}\|^2=1$.

Proof. According to the relation (7) where $i=(\ell-1)N+j$ with $(1\le\ell\le P)$ and $(1\le j\le N)$, the theorem results in Theorem 1, part 1. □

In the mixed subband scheme, the matrix $\mathbf{Y}$ is of dimension $NP\times K$. If we apply one quantizer per element of blocks located in a subband of a component (in other words one quantizer per row of $\mathbf{Y}$ and per subband) and if we suppose high-rate quantizations (see Appendix A), then condition $\mathcal{H}_1$ is approximatively satisfied and the distortion $D_{j,\ell,k}$ does not depend on the spatial localization of the block k in the subband of the component [6,28]. As a consequence we obtain the following corollary.

Corollary 1. 1. *With the notations of Section* 2.3.2, *with one quantizer per element of blocks located in a subband of a component* (*i.e., one quantizer per row of* $\mathbf{Y}$ *and per subband*) *and under the assumptions* $\mathcal{H}_2$, $\mathcal{H}_3$ *and high resolution quantization hypotheses, the end-to-end distortion becomes*

$$D=\frac{1}{NP}\sum_{j=1}^{N}\sum_{\ell=1}^{P}\sum_{m=1}^{M}\pi_m\omega_m w_{j,\ell}^{(m)}D_{j,\ell}^{(m)}, \quad (9)$$

with $\pi_m=K_m/K=PK_m/L$ *the proportion of wavelet coefficients in the subband m,* $\omega_m=(1/K_m)\sum_k\|\mathbf{W}^{-1}\mathbf{e}_{kP}\|^2$ (*in the summation the range of k consists of the* K_m *indexes of columns of* $\mathbf{Y}$ *associated with the subband m*), $w_{j,\ell}^{(m)}=\|\mathbf{A}^{(m)^{-1}}\mathbf{e}_{(\ell-1)N+j}\|^2$ *and* $D_{j,\ell}^{(m)}$ *is the mean distortion of the* ℓ*th element* $(1\le\ell\le P)$ *of a block in the subband m of the jth component.*

2. *If the DWT* $\mathbf{W}$ *and the spectral transforms* $\mathbf{A}^{(m)}$ *are all orthogonal, then the relation* (9) *remains valid without the assumption* $\mathcal{H}_1$ *and in this case we have* $\omega_m=w_{j,\ell}^{(m)}=1$ *for all j,* ℓ *and m.*

Remark 2. The weightings $(\omega_m)_{1\le m\le M}$ depend only on the synthesis filter of the DWT. The weighting $w_{j,\ell}^{(m)}$ depends only on the spectral transform applied to the subband m.

In the particular case where all the $\mathbf{A}^{(m)}$ are orthogonal, the distortion expression reduces to

$$D=\frac{1}{NP}\sum_{j=1}^{N}\sum_{\ell=1}^{P}\sum_{m=1}^{M}\pi_m\omega_m D_{j,\ell}^{(m)},$$

since the coefficients $w_{j,\ell}^{(m)}$ are then all equal to 1.

3.3. The subband and separable schemes

The subband scheme is a particular case of the mixed subband scheme with $P=1$, therefore the relation (5) leads to $\|\mathcal{A}^{-1}\mathrm{vec}(\mathbf{e}_{ik})\|^2=\|\mathbf{A}^{(m)^{-1}}\mathbf{e}_i\|^2\|\mathbf{W}^{-1}\mathbf{e}_k\|^2$, which can be used to prove the next corollary, following the same reasoning as in Section 3.2.

Corollary 2. 1. *Under the assumption* $\mathcal{H}_1$, *with one quantizer per subband and per component and with the notations of Section* 2.3.1, *the end-to-end distortion of*

the subband scheme is given by

$$D=\frac{1}{N}\sum_{m=1}^{M}\pi_m\omega_m\left[\sum_{i=1}^{N}w_i^{(m)}D_i^{(m)}\right], \tag{10}$$

where π_m is the proportion of wavelet coefficients in the subband m, $\omega_m=(1/K_m)\sum_k\|\mathbf{W}^{-1}\mathbf{e}_k\|^2$ (in the summation the range of k consists of the K_m indexes of columns of $\mathbf{Y}$ associated with the subband m), $w_i^{(m)}=\|\mathbf{A}^{(m)^{-1}}\mathbf{e}_i\|^2$ and $D_i^{(m)}$ is the quantizer distortion of the subband m of the ith component.

2. *If the DWT $\mathbf{W}$ and the spectral transforms $\mathbf{A}^{(m)}$ are all orthogonal, then the relation* (10) *remains valid without the assumption $\mathcal{H}_1$ and in this case we have $\omega_m=w_i^{(m)}=1$ for all i and m.*

The expression between the brackets can be viewed as the distortion of the subband m in the wavelet domain. The assumption $\mathcal{H}_2$ is not required for the subband scheme distortion formula (10), it suffices to assume the condition $\mathcal{H}_1$ in the part 1 of Corollary 2. Finally, for the separable scheme where the same spectral transform is applied to all subbands, the distortion formula differs only in the weightings $w_i^{(m)}$ which do not have to depend on m anymore. Then the following corollary holds.

Corollary 3. 1. *Under the assumption $\mathcal{H}_1$, with one quantizer per subband and per component and with the notations of Section* 2.2, *the end-to-end distortion of the separable scheme is given by*:

$$D=\frac{1}{N}\sum_{m=1}^{M}\pi_m\omega_m\left[\sum_{i=1}^{N}w_iD_i^{(m)}\right], \tag{11}$$

where π_m is the proportion of wavelet coefficients in the subband m, $\omega_m=(1/K_m)\sum_k\|\mathbf{W}^{-1}\mathbf{e}_k\|^2$ (in the summation, the range of k consists of the columns of $\mathbf{Y}$ associated with the subband m), $w_i=\|\mathbf{A}^{-1}\mathbf{e}_i\|^2$ and $D_i^{(m)}$ is the quantizer distortion of the subband m of the ith component.

2. *If both the DWT $\mathbf{W}$ and the spectral transform $\mathbf{A}$ are orthogonal, then the relation* (11) *remains valid without the assumption $\mathcal{H}_1$ and in this case we have $\omega_m=w_i=1$ for all i and m.*

Remark 3. The assumption $\mathcal{H}_1$ is a consequence of high resolution quantizations (see Appendix A). It can also be deduced from the condition of statistical independence of the transformed components, since if the components of $\mathbf{Y}$ are independent, then the components of the quantization noise $\mathbf{Y}-\mathbf{Y}^q$ are uncorrelated.

A method for the computation of the weighting wavelet coefficients ω_m ($1\le m\le M$) can be found in [27,28]. Since the Daubechies 9/7 DWT is only quasi-orthogonal, the equalities (8), (9) and (11) are actually good approximations at medium bit-rates. We show in Section 6.2 the accuracy of these approximations.

For each compression scheme, we search the optimal spectral transform (that is the one which minimizes the total bit-rate for a given end-to-end distortion) which adapts to the data, assuming high resolution quantization hypotheses and 2-D DWT with fixed coefficients, i.e., which do not adapt to the data. As already mentioned, in our tests we always used the Daubechies 9/7 DWT. First, we derive the criterion minimized by an optimal spectral transform. We emphasize the fact that we do not assume Gaussian data and that generally in the literature this assumption is made in order to clarify the criterion (coding gain) maximized by the optimal transform. However, Bennett's formula and the optimal bit allocation between quantizers formula on which our criteria are based are well-known and therefore it is straightforward to deduce these criteria from well-known results. Our major innovation consists especially in the computation of the optimal transforms, since this computation is generally presented as a difficult task [19] in classical transform coding and has never been done in the case of the separable scheme which is JPEG2000 compatible. The computation of the optimal transforms in both cases of subband and mixed subband schemes is the same as in transform coding of still images and has been presented in [20].

4. Criteria for optimal transforms under high resolution quantizations

We recall the extension of Bennett's formula which can be stated as follows: if X is a real random variable quantized under the high resolution hypothesis, then the bit-rate of quantized variable X^q is well approximated by $H(X)-\frac{1}{2}\log_2(cD)$, where $H(X)$ is the differential entropy of X, D is the distortion (expected mean square error) introduced by the quantization and c is a constant depending on the quantization, e.g., for uniform scalar quantization $c=12$ [29].

Theorem 3 (*Mixed subband scheme*). *For the mixed subband scheme when the 2-D DWT has fixed coefficients, under the high resolution quantization hypotheses and the conditions $\mathcal{H}_2$ and $\mathcal{H}_3$, the optimal spectral transform $\mathbf{A}^{(m)}$ associated with the subband m is an $NP\times NP$ matrix that minimizes the criterion*

$$C_1(\mathbf{A}^{(m)})=\sum_{j=1}^{N}\sum_{\ell=1}^{P}H(Y_{j,\ell}^{(m)})+\frac{1}{2}\log_2\det\operatorname{diag}[\mathbf{A}^{(m)^{-T}}\mathbf{A}^{(m)^{-1}}]. \tag{12}$$

Proof. From Corollary 1, we know that the distortion is given by Eq. (9). Moreover, if $R_{j,\ell}^{(m)}$ denotes the bit-rate of the quantizer associated with the ℓth element of blocks of dimension P located in the subband m of the jth component, then it is clear that the total bit-rate satisfies

$$R=\frac{1}{NP}\sum_{j=1}^{N}\sum_{m=1}^{M}\sum_{\ell=1}^{P}\pi_mR_{j,\ell}^{(m)}. \tag{13}$$

Therefore, according to Bennett's approximation, the total bit-rate verifies

$$R\simeq\frac{1}{NP}\sum_{j=1}^{N}\sum_{\ell=1}^{P}\sum_{m=1}^{M}\pi_m\left[H(Y_{j,\ell}^{(m)})-\frac{1}{2}\log_2(cD_{j,\ell}^{(m)})\right] \tag{14}$$

and the problem now consists in minimizing the total bit-rate for a given maximal end-to-end distortion D_t. It is a

classical problem in compression, called optimal bit allocation, and it can be proven [26] (see Appendix B) that the minimization holds when $D_{j,\ell,opt}^{(m)} = D_t/\omega_m w_{j,\ell}^{(m)}$ for all ℓ, m and j. Hence, the optimal bit allocation leads to the next approximation between the total bit-rate R and the end-to-end distortion D_t

$$R \simeq \sum_{m=1}^{M} \pi_m \left[\frac{1}{NP} \sum_{j=1}^{N} \sum_{\ell=1}^{P} \left\{ H(Y_{j,\ell}^{(m)}) + \frac{1}{2}\log_2 w_{j,\ell}^{(m)} \right\} + \frac{1}{2}\log_2 w_m \right] - \frac{1}{2}\log_2(cD_t). \quad (15)$$

Since the 2-D DWT has fixed coefficients, the weightings w_m are fixed and the optimal spectral transform associated with the subband m minimizes $\sum_{j=1}^{N}\sum_{\ell=1}^{P}[H(Y_{j,\ell}^{(m)}) + \frac{1}{2}\log_2(w_{j,\ell}^{(m)})]$ where—according to Corollary 1 and relation (3)—$w_{j,\ell}^{(m)} = \sum_{i=1}^{NP} (\mathbf{A}^{(m)^{-1}})_{ij}^2$. We recall that for the mixed subband scheme the spectral transform $\mathbf{A}^{(m)}$ is an $NP \times NP$ matrix. To end the proof, we remark that the weighting $w_{j,\ell}^{(m)}$ is equal to the $[P(j-1)+\ell]$th element of $\mathrm{diag}(\mathbf{A}^{(m)^{-T}}\mathbf{A}^{(m)^{-1}})$. □

Corollary 4 (*Subband scheme*). *For the subband scheme when the 2-D DWT has fixed coefficients, under the only high resolution quantization hypotheses, the optimal spectral transform* $\mathbf{A}^{(m)}$ *associated with the subband* m *is an* $N \times N$ *matrix that minimizes the criterion*

$$C_1(\mathbf{A}^{(m)}) = \sum_{j=1}^{N} H(Y_j^{(m)}) + \frac{1}{2}\log_2 \det \mathrm{diag}[\mathbf{A}^{(m)^{-T}}\mathbf{A}^{(m)^{-1}}].$$

Proof. It is a straightforward consequence of both Corollary 2 and Theorem 3. □

We can notice that for both the mixed subband scheme and the subband scheme, the criterion $C_1(\mathbf{A}^{(m)})$, where the exponent $^{(m)}$ has been removed to simplify the notations, can be rewritten—like in [20]—as: $C_1(\mathbf{A}) = C_{ICA}(\mathbf{A}) + C_O(\mathbf{A})$, with

$$C_O(\mathbf{A}) = \frac{1}{2}\log_2 \frac{\det \mathrm{diag}(\mathbf{A}^{-T}\mathbf{A}^{-1})}{\det(\mathbf{A}^{-T}\mathbf{A}^{-1})}$$

and $C_{ICA}(\mathbf{A}) = \sum_{i=1}^{N} H(Y_i) - \log_2 \det\mathbf{A}$ is the criterion to minimize when performing only ICA. Pham [24] used the criterion $C_{ICA}(\mathbf{A})$ to perform the algorithm `ICAinf` that is based on the minimization of the mutual information of the components of $\mathbf{Y}$.

Remark 4. The term $C_O(\mathbf{A})$ is always positive or null [20], it vanishes if and only if $\mathbf{A}$ is a matrix whose columns are pairwise orthogonal, therefore it can be seen like a kind of measure of deviation to orthogonality.

Theorem 4 (*Separable scheme*). *For the separable scheme when the 2-D DWT has fixed coefficients, if high resolution quantization hypotheses are assumed, then the optimal spectral transform* $\mathbf{A}$ *is an* $N \times N$ *matrix that minimizes the criterion*:

$$C_2(\mathbf{A}) = \sum_{j=1}^{N}\sum_{m=1}^{M} \pi_m H(Y_j^{(m)}) + \frac{1}{2}\log_2 \det \mathrm{diag}(\mathbf{A}^{-T}\mathbf{A}^{-1}). \quad (16)$$

Proof. For a total bit-rate R and a maximal end-to-end distortion D_t, the approximation (14) becomes

$$R \simeq \frac{1}{N}\sum_{j=1}^{N}\sum_{m=1}^{M} \pi_m \left[H(Y_j^{(m)}) - \frac{1}{2}\log_2(cD_j^{(m)}) \right], \quad (17)$$

constrained to the condition (according to Corollary 3) $(1/N)\sum_{m=1}^{M}\sum_{i=1}^{N} \pi_m \omega_m w_j D_j^{(m)} \le D_t$. Then, the same reasoning as that given in Appendix B shows that the optimal bit allocation is achieved when the distortions in all subbands satisfy $D_{j,opt}^{(m)} = D_t/\omega_m w_j$ with w_j the jth diagonal element of $\mathbf{A}^{-T}\mathbf{A}^{-1}$. The introduction of the optimal values in the expression (17) completes the proof. □

Remark 5. Since $\sum_{m=1}^{M} \pi_m = 1$, the criterion $C_2(\mathbf{A})$ can be expressed as

$$\begin{aligned} C_2(\mathbf{A}) &= \sum_{m=1}^{M} \pi_m \left[\sum_{i=1}^{N} H(Y_i^{(m)}) - \log_2|\det\mathbf{A}| \right] + \frac{1}{2}\log_2 \left[\frac{\det \mathrm{diag}(\mathbf{A}^{-T}\mathbf{A}^{-1})}{\det(\mathbf{A}^{-T}\mathbf{A}^{-1})} \right] \\ &= \sum_{m=1}^{M} \pi_m C_{ICA}^{(m)}(\mathbf{A}) + C_O(\mathbf{A}), \end{aligned} \quad (18)$$

where, for $1 \le m \le M$, $C_{ICA}^{(m)}(\mathbf{A}) = \sum_{i=1}^{N} H(Y_i^{(m)}) - \log_2|\det\mathbf{A}|$ is the criterion of ICA applied to the N components of the transformed coefficients that belong to the subband m.

The relation (18) shows that the criterion $C_2(\mathbf{A})$ of the separable scheme takes into consideration the fact that one quantizer per subband and per component is allocated. It is also important to notice that the criterion $C_2(\mathbf{A})$ involves the transformed coefficients $\mathbf{Y}$. Therefore, even for the separable scheme (where the order of processing between the 2-D DWT and the spectral transform does not matter), the search of the optimal spectral transform must be done *after* the 2-D DWT.

5. Minimization of the criteria for the separable scheme

For both the mixed subband scheme and the subband scheme, the algorithm that minimizes the criterion $C_1(\mathbf{A})$ is called `GCGsup`. Another one that minimizes $C_1(\mathbf{A})$ constrained to the condition that $C_O(\mathbf{A}) = 0$ is called `OrthICA`. The detailed description of these algorithms can be found in [20] or [30].

For the separable scheme, we explain now two algorithms that minimize the criterion (16), one without no constraint but invertibility and the other with the constraint of orthogonality. To simplify some mathematical expressions we shall use the Neperian logarithm instead of the base two logarithm until the end of this section. As in [24,20], the algorithms of minimization are based on a quasi-Newton method with the relative gradient and a simplified relative Hessian. Starting with a current estimator $\mathbf{A}$, the method consists of expanding $C_2(\mathbf{A} + \mathcal{E}\mathbf{A})$ with respect to the matrix $\mathcal{E} = [\mathcal{E}_{ij}]$ up to the second order, in a neighborhood of $\mathcal{E} = \mathbf{0}_N$ (the null matrix), and then minimizing the resulting quadratic form in $\mathcal{E}$ to obtain a new estimate. Using the results of [31] it is straightforward to deduce that the Taylor

expansion up to the second order of $C_{ICA}^{(m)}(\mathbf{A}+\mathcal{E}\mathbf{A})$ can be approximated as follows:

$$C_{ICA}^{(m)}(\mathbf{A}+\mathcal{E}\mathbf{A}) = C_{ICA}^{(m)}(\mathbf{A}) + \sum_{1\le i\ne j\le N} E[\psi_{Y_i^{(m)}}(Y_i^{(m)})Y_j^{(m)}]\mathcal{E}_{ij} + \frac{1}{2}\sum_{1\le i\ne j\le N}\{E[\psi^2_{Y_i^{(m)}}(Y_i^{(m)})]E[Y_j^{(m)2}]\mathcal{E}_{ij}^2 + \mathcal{E}_{ij}\mathcal{E}_{ji}\} + \cdots, \tag{19}$$

where the function $\psi_{Y_i^{(m)}}$ is equal to the derivative of $-\log p(y_i^{(m)})$—$p(y_i^{(m)})$ denoting the probability density function of $Y_i^{(m)}$—and is known as the score function. Let $\mathbf{M}=\mathbf{A}^{-T}\mathbf{A}^{-1}$. In [20], the Taylor expansion of $C_O(\mathbf{A}+\mathcal{E}\mathbf{A})$ is given up to the second order; however, it is quite involved and it is simplified into

$$C_O(\mathbf{A}+\mathcal{E}\mathbf{A}) \approx C_O(\mathbf{A}) - \sum_{1\le i\ne j\le N}\frac{M_{ji}}{M_{ii}}\mathcal{E}_{ji} + \frac{1}{2}\sum_{1\le i\ne j\le N}\left[\frac{M_{jj}}{M_{ii}}\mathcal{E}_{ji}^2 + \mathcal{E}_{ji}\mathcal{E}_{ij}\right] + \cdots \tag{20}$$

by neglecting the non-diagonal elements of $\mathbf{M}=[M_{ij}]$ in the second order terms of the Taylor expansion.

Using the approximation (20), the equality (19) and the relation (18) we obtain

$$C_2(\mathbf{A}+\mathcal{E}\mathbf{A}) = C_2(\mathbf{A}) + \sum_{1\le i\ne j\le N}\left[\sum_{m=1}^{M}\pi_m E[Y_j^{(m)}\psi_{Y_i^{(m)}}(Y_i^{(m)})] - \frac{\mathbf{M}_{ij}}{\mathbf{M}_{jj}}\right]\mathcal{E}_{ij} + \frac{1}{2}\sum_{1\le i\ne j\le N}\left[\sum_{m=1}^{M}\pi_m\mathcal{E}_{ij}^2 E[Y_j^{(m)2}]E[\psi^2_{Y_i^{(m)}}(Y_i^{(m)})] + \frac{M_{ii}}{M_{jj}}\mathcal{E}_{ij}^2 + 2\mathcal{E}_{ij}\mathcal{E}_{ji}\right]. \tag{21}$$

The quadratic form associated to this last expansion is positive definite. One iteration of the algorithm is first to solve the following equation:

$$\begin{bmatrix}\Psi_{ij} & 2\\ 2 & \Psi_{ji}\end{bmatrix}\begin{pmatrix}\mathcal{E}_{ij}\\ \mathcal{E}_{ji}\end{pmatrix} = \begin{pmatrix}\Phi_{ij}\\ \Phi_{ji}\end{pmatrix}, \tag{22}$$

with $\Phi_{ij}=M_{ij}/M_{jj}-\sum_{m=1}^{M}\pi_m E[\psi_{Y_i^{(m)}}(Y_i^{(m)})Y_j^{(m)}]$ and $\Psi_{ij}=\sum_{m=1}^{M}\pi_m E[\psi^2_{Y_i^{(m)}}(Y_i^{(m)})]E[Y_i^{(m)2}]+M_{ii}/M_{jj}$ and then to replace the current solution $\mathbf{A}$ with $\mathbf{A}+\mathcal{E}\mathbf{A}$. Since the diagonal elements of $\mathcal{E}$ are undetermined, they are arbitrarily fixed to zero. For the practical computation of the algorithm, we replace $\psi_{Y_i^{(m)}}$ with its estimator $\hat{\psi}_{Y_i^{(m)}}$ that is described in [31] as well as the estimator of the differential entropy. The mathematical expectations are replaced with simple empirical means. We call `OST` (*optimal spectral transform*) the algorithm described above and also the optimal transform returned by this algorithm.

To minimize the criterion (16) with the constraint that the solution is an orthogonal matrix, it is important to note, as in [20], that if $\mathbf{A}$ is orthogonal, then $\mathbf{A}+\mathcal{E}\mathbf{A}$ remains orthogonal when $\mathbf{I}+\mathcal{E}$ is also orthogonal. This condition is satisfied up to the first order if $\mathcal{E}$ is an antisymmetrical matrix, since then $(\mathbf{I}+\mathcal{E})^T\ (\mathbf{I}+\mathcal{E})=\mathbf{I}+\mathcal{E}^T\mathcal{E}$. Using that condition, the expansion (21) becomes

$$C(\mathbf{A}+\mathcal{E}\mathbf{A}) = C(\mathbf{A}) + \sum_{m=1}^{M}\sum_{1\le i<j\le N}\pi_m\{E[Y_j^{(m)}\psi_{Y_i^{(m)}}(Y_i^{(m)})] - E[Y_i^{(m)}\psi_{Y_j^{(m)}}(Y_j^{(m)})]\}\mathcal{E}_{ij} + \frac{1}{2}\sum_{1\le i<j\le N}\left[\sum_{m=1}^{M}\pi_m\{E[Y_j^{(m)2}]E[\psi^2_{Y_i^{(m)}}(Y_i^{(m)})] + E[Y_i^{(m)2}]E[\psi^2_{Y_j^{(m)}}(Y_j^{(m)})]\} - 2\right]\mathcal{E}_{ij}^2. \tag{23}$$

The matrix $\mathcal{E}$ is calculated in that case according to

$$\mathcal{E}_{ij} = \frac{\sum_{m=1}^{M}\pi_m\{E[Y_i^{(m)}\psi_{Y_j^{(m)}}(Y_j^{(m)})] - E[Y_j^{(m)}\psi_{Y_i^{(m)}}(Y_i^{(m)})]\}}{\sum_{m=1}^{M}\pi_m\{E[Y_j^{(m)2}]E[\psi^2_{Y_i^{(m)}}(Y_i^{(m)})] + E[Y_i^{(m)2}]E[\psi^2_{Y_j^{(m)}}(Y_j^{(m)})]\} - 2}. \tag{24}$$

Actually, $\mathbf{A}+\mathcal{E}\mathbf{A}$ obtained in this way is not a true orthogonal matrix. This can be overcome by replacing $\mathbf{A}+\mathcal{E}\mathbf{A}$ with $e^{\mathcal{E}}\mathbf{A}=(\mathbf{I}+\mathcal{E}+\mathcal{E}^2/2!+\cdots)\mathbf{A}$, which is orthogonal and differs from $\mathbf{A}+\mathcal{E}\mathbf{A}$ only by second order terms. We call both `OrthOST` (*orthogonal optimal spectral transform*) this algorithm and the orthogonal transform returned by the algorithm. The case where the spectral transform is constrained to be orthogonal is particularly interesting because the weightings which depend on the linear transform are all equal to one.

5.1. Computational complexity of the optimal transforms

We give here a rough estimation of the number of operations required for the computation of both algorithms described above, taking into account only multiplications and divisions. The differential entropies and the score functions are calculated according to a method explained in [31]. The computational complexity of each of these quantities is $O(NrL)$, where r is the number of bins in the binned kernel density estimation. In general $r\ll L$ and for most cases r belongs to the interval [30,60]. At each iteration, the criterion and the matrix $\mathcal{E}$ must be computed. The complexity of the criterion computation is $O(NrL+N^3)$. For the calculation of the matrix $\mathcal{E}$, we first need to compute the score function. The complexity of the matrix $\mathcal{E}$ computation (including the score function computation) is $O(NrL+N^2L)$. Finally, the complexity of one iteration is $O(NrL+N^2L+N^3)$. In practice, the convergence of the algorithm is usually obtained after p iterations, $p\in[20,60]$. Generally $N\ll L$ and the total computational complexity is $O(p(NrL+N^2L))$. The computational complexities of `OrthOST` and of both algorithms `GCGsup` and `OrthICA` are the same. We recall that for the KLT, this complexity is $O(LN^2)$.

6. Experimental results

In this section we present the performances in image compression of the optimal transforms described in the previous section.

6.1. Description of the tests

As already mentioned, the 2-D DWT used in all our experiments is the Daubechies 9/7 which proved to be efficient in lossy image compression [4,6]. For simplicity, we used only uniform scalar quantizers with a dead zone twice as large as the quantization step. The performances are evaluated in terms of bit-rate versus end-to-end distortion. For hyperspectral images, we considered four distortions. A first one is the mean square error (MSE) expressed in terms of the SNR (signal to noise ratio), $\mathrm{SNR} = 10\log_{10}\sigma^2/D$ where D is the actual end-to-end MSE distortion and σ^2 is the empirical variance of the initial image: $\sigma^2 = (1/NL)\sum_{i=1}^{N}\sum_{n=1}^{L}(X_i(n)-\mu)^2$, with $\mu = (1/NL)\sum_{i=1}^{N}\sum_{n=1}^{L}X_i(n)$ the empirical mean of the image. A second distortion is the maximal absolute difference ($\mathrm{MAD} = \max\{|X_i(n)-\hat{X}_i(n)|: 1\le i\le N \text{ and } 1\le n\le L\}$), a third one is the maximum spectral angle

$$\mathrm{MSA} = \max\left\{\operatorname{acos}\left(\frac{\sum_{i=1}^{N}X_i(n)\hat{X}_i(n)}{\sqrt{\sum_{i=1}^{N}X_i^2(n)\sum_{i=1}^{N}\hat{X}_i^2(n)}}\right): 1\le n\le L\right\} \tag{25}$$

and the last one is the mean absolute error ($\mathrm{MAE} = (1/NL)\sum_{i=1}^{N}\sum_{n=1}^{L}|X_i(n)-\hat{X}_i(n)|$). With these four distortions, one can estimate the performances of a codec on usual applications of hyperspectral images, like classifications and targets detections [25]. For multispectral images, we considered only the MAD and the MSE distortions, the last one being expressed in terms of PSNR (peak of signal to noise ratio): $\mathrm{PSNR} = 10\log_{10}(2^{N_b}-1)^2/D$, where D is the actual end-to-end MSE distortion and N_b is the number of bits per pixel and per band (bpppb) of the initial image. The bit-rate, expressed in bpppb, was either estimated by the first order entropy (more precisely by an average of the first order entropy per subband and per component weighted by the ratio of the subband size) or measured on the actual bit stream obtained with the JPEG2000 coder EBCOT [7] and its PCRD optimizer applied across components for optimal bit allocation. In this case, we used the Verification Model version 9.1 (VM9 [32]) codec developed by the JPEG2000 group and we applied it only to the separable scheme. When the VM9 is used, the coefficients of $\mathbf{A}^{-1}$ (the inverse matrix of the optimal spectral transform) and the mean of each component are stored in the bitstream as `float32` data (this costs $32(N+1)/L$ bpppb). When the average first order entropy is used to estimate the bit-rate, the optimal bit allocation between subbands and components is done with the well known algorithm by Shoham and Gersho [33] and the size of the bit stream occupied by the inverse transforms ($\mathbf{A}^{-1}$ or $(\mathbf{A}^{(m)^{-1}})_m$) and by the means of the components is not taken into account.

We tested two kinds of multicomponent images: multispectral ones and hyperspectral ones. The multispectral images are[1] PLEIADES simulations of French cities with $N = 4$ components and coded on $N_b = 12$ bpppb: *Moissac* with $N_c \times N_r = 320 \times 3152$, *Port-de-Bouc* with $N_c \times N_r = 320 \times 1376$, *Toulouse* with $N_c \times N_r = 352 \times 3816$, *Vannes* with $N_c \times N_r = 352 \times 3736, \ldots$. The hyperspectral images are[2] AVIRIS images (*Moffett*, *Cuprite* and *Jasper*) with $N = 224$ components from the visible to the infrared and coded on $N_b = 16$ bpppb. They are originally acquired with $N_r \times N_c = 512 \times 624$, but for the simulations we kept only the 512 leftmost columns. Some images used in our tests are shown in Appendix C.

6.2. Validation of the distortion formulae

Before presenting the performances of the optimal transforms, let us recall that their computation is strongly based on the distortion formulae given in Section 3, which are valid under the assumptions $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$. In our experiments these assumptions are only approximatively satisfied, therefore the distortion equalities would be only approximations that have to be validated.

For this purpose, we randomly distributed the quantization steps in each block of each subband, then we computed the end-to-end distortion in two different ways. One consists in applying the distortion formulae of (9) to obtain the end-to-end distortion (denoted D_{app}) from the quantizer distortions $D_{j,\ell}^{(m)}$ ($1\le j\le N$, $1\le \ell\le P$, $1\le m\le M$), which are evaluated in the transform domain. The other consists in computing the actual end-to-end distortion (denoted D_r) from the reconstructed image. For the three compression schemes and for all the transforms tested, we observed that for distortions associated with bit-rates greater than 0.25 bpppb the ratio D_{app}/D_r belongs to [0.98,1]. In other words, the distortion formulae are good approximations at medium and high bit-rates (greater than 0.25 bpppb).

6.3. Bit-rate versus distortion performances

In this subsection, we discuss and compare the bit-rate versus distortion performances of different spectral transforms for the three compression schemes.

For the separable scheme, Table 1 presents the bit-rate of different transforms versus the two distortions PSNR and MAD on six multispectral images and Tables 2 and 3 present the bit-rate of different transforms versus the four distortions SNR (in dB), MAE, MAD and MSA (expressed in degree (°)) on three hyperspectral images. The bit-rates was computed with the VM9 (hence all the information required for decoding was taken into account) and all the 2-D DWT was applied with five levels of decomposition.

We observe the well-known fact that spectral transforms perform significantly better than the identity matrix (i.e., no spectral transform), especially for hyperspectral images. Indeed, on six multispectral images (see Table 1) the average gains of the KLT, `OrthOST` and `OST` on Identity are, respectively, 4.2, 4.5 and 4.5 dB. On three hyperspectral images (see Table 2) the average gains of the KLT, `OrthOST` and `OST` on Identity are, respectively, 15.9,

[1] These images have been given by the French Space Agency CNES (Centre National d'Etudes Spatiales). They are described on the web site http://smsc.cnes.fr/PLEIADES/.

[2] These images have been downloaded from the NASA web site http://aviris.jpl.nasa.gov/.

Table 1
Bit-rate (in bpppb) versus PSNR (in dB) and versus MAD of different spectral transforms on multispectral images for the separable scheme (best results are bolded).

Bit-rate	PSNR (dB)								MAD							
	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
Moissac																
Id	36.37	39.59	41.93	43.89	47.22	50.16	52.93	55.66	**691**	**366**	253	187	108	68	49	38
KLT	38.61	42.39	45.24	47.63	51.51	54.49	56.98	59.44	716	381	**214**	**135**	79	48	**32**	25
OrthOST	38.67	42.50	45.35	47.72	51.55	54.55	57.11	59.60	818	399	229	145	**78**	**47**	33	24
OST	**38.69**	**42.55**	**45.43**	**47.80**	**51.62**	**54.59**	**57.15**	**59.65**	745	496	215	138	**78**	48	**32**	**23**
Port-de-Bouc																
Id	30.36	33.68	36.14	38.25	41.93	45.27	48.43	51.52	1198	653	544	361	198	135	85	64
KLT	**33.47**	37.74	40.88	43.45	47.53	50.89	53.82	56.53	922	**513**	**297**	**230**	139	**74**	**50**	35
OrthOST	33.42	37.80	41.05	43.71	47.90	51.31	54.28	56.99	885	**513**	305	237	**122**	77	51	**31**
OST	33.46	**37.85**	**41.12**	**43.78**	**48.00**	**51.40**	**54.36**	**57.06**	**866**	557	351	256	129	82	53	35
Vannes																
Id	39.25	42.89	45.67	47.99	51.77	54.80	57.51	60.11	603	269	178	109	63	42	29	21
KLT	41.36	45.71	48.78	51.11	54.38	56.82	59.24	61.79	482	219	148	**86**	51	33	**24**	18
OrthOST	41.90	46.27	49.29	51.54	54.71	57.18	59.62	62.16	**354**	**190**	**135**	91	46	**30**	25	18
OST	**41.94**	**46.34**	**49.35**	**51.59**	**54.74**	**57.22**	**59.68**	**62.20**	393	204	138	88	**45**	33	25	**16**
Strasbourg																
Id	30.82	34.19	36.73	38.91	42.70	46.09	49.20	52.13	1357	**877**	546	353	205	118	86	60
KLT	**32.51**	**36.59**	39.77	42.49	46.99	50.58	53.51	56.08	1041	927	**438**	403	184	90	52	**38**
OrthOST	**32.51**	**36.59**	39.78	42.50	47.01	50.61	53.55	56.11	**1010**	948	449	404	178	87	**50**	**38**
OST	32.49	**36.59**	**39.79**	**42.53**	**47.07**	**50.67**	**53.60**	**56.17**	1149	904	455	**289**	**162**	**81**	55	42
Montpellier																
Id	32.17	35.23	37.59	39.62	43.17	46.30	49.17	51.95	1216	630	406	292	168	117	77	54
KLT	34.09	37.75	40.60	43.03	47.20	50.69	53.63	56.18	747	488	340	248	143	75	49	34
OrthOST	34.08	37.90	40.92	43.46	47.72	51.16	54.01	56.55	**681**	**454**	338	255	**127**	**68**	47	**32**
OST	**34.14**	**37.99**	**41.01**	**43.56**	**47.79**	**51.21**	**54.06**	**56.60**	704	483	**332**	**239**	**127**	**68**	**46**	33
Perpignan																
Id	33.71	36.90	39.34	41.43	45.04	48.17	51.04	53.78	984	526	332	230	158	89	62	42
KLT	36.51	40.44	43.29	45.60	49.33	52.36	54.99	57.52	726	435	245	172	**84**	**54**	41	29
OrthOST	36.59	40.59	43.48	45.83	49.61	52.66	55.30	57.82	721	**371**	**232**	165	94	**54**	**37**	30
OST	**36.60**	**40.60**	**43.49**	**45.84**	**49.62**	**52.67**	**55.32**	**57.85**	**645**	383	292	**164**	94	55	38	**28**

The bit-rate was computed with the VM9.

Table 2
Bit-rate (in bpppb) versus SNR (in dB) and versus MAE of different spectral transforms on hyperspectral images for the separable scheme.

Bit-rate	SNR (dB)								MAE							
	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
Moffett																
Id	25.45	30.37	33.97	36.94	41.78	45.76	49.15	52.01	39.57	24.32	16.93	12.52	7.62	5.02	3.48	2.51
KLT	44.21	47.68	50.08	51.97	54.76	57.10	59.21	61.04	5.36	3.83	3.03	2.49	1.82	1.39	1.07	0.85
OrthOST	45.31	**48.57**	**50.87**	52.61	55.28	57.57	59.62	61.37	4.77	**3.47**	**2.78**	**2.32**	**1.72**	1.31	1.02	0.81
OST	**45.32**	48.56	**50.87**	**52.62**	**55.30**	**57.64**	**59.77**	**61.63**	**4.75**	**3.47**	**2.78**	**2.32**	**1.72**	**1.30**	**1.00**	**0.79**
Cuprite																
Id	29.99	33.48	36.12	38.41	42.44	45.99	49.19	52.11	37.14	26.07	19.85	15.60	10.13	6.89	4.83	3.47
KLT	47.79	50.46	52.55	54.16	56.76	59.07	61.26	63.27	5.11	3.96	3.23	2.73	2.04	1.55	1.19	0.92
OrthOST	48.25	50.88	**52.89**	**54.44**	56.99	59.29	61.46	63.44	**4.83**	**3.79**	**3.12**	**2.65**	**1.98**	1.51	1.16	0.90
OST	**48.26**	**50.89**	**52.89**	**54.44**	**57.01**	**59.34**	**61.56**	**63.60**	**4.83**	**3.79**	**3.12**	**2.65**	**1.98**	**1.50**	**1.14**	**0.88**
Jasper																
Id	21.34	24.83	27.56	29.92	34.01	37.67	41.09	44.33	64.84	45.61	34.39	26.82	17.23	11.52	7.89	5.49
KLT	42.93	46.49	48.61	50.37	53.18	55.56	57.72	59.66	5.78	4.04	3.27	2.72	1.99	1.51	1.16	0.91
OrthOST	43.66	46.94	49.02	50.73	53.47	55.81	57.94	59.85	5.35	3.85	3.13	2.62	1.93	1.46	1.13	0.88
OST	**43.70**	**46.96**	**49.05**	**50.74**	**53.50**	**55.87**	**58.03**	**60.01**	**5.32**	**3.84**	**3.12**	**2.61**	**1.92**	**1.45**	**1.11**	**0.87**

The bit-rate was computed with the VM9.

Table 3
Bit-rate (in bpppb) versus MSA (in degree (°)) and versus MAD of different spectral transforms on hyperspectral images for the separable scheme.

Bit-rate	MSA (deg)								MAD							
	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
Moffett																
Id	12.12	6.82	3.94	2.66	1.29	0.85	0.52	0.36	1676	781	492	1259	183	62	32	20
KLT	1.43	0.87	0.57	0.37	0.20	0.15	0.12	0.10	392	211	119	67	24	14	8	7
OrthOST	0.96	**0.47**	**0.31**	**0.25**	**0.18**	**0.14**	**0.11**	**0.09**	261	**77**	49	**33**	**18**	**10**	8	**6**
OST	**0.86**	0.50	0.32	**0.25**	**0.18**	**0.14**	**0.11**	**0.09**	**207**	101	**46**	37	19	12	**7**	**6**
Cuprite																
Id	5.30	2.81	2.20	1.57	1.01	0.59	0.40	0.26	659	360	253	185	110	62	61	40
KLT	0.42	0.25	0.22	0.15	0.12	**0.08**	**0.07**	0.06	154	135	100	54	26	16	10	8
OrthOST	**0.32**	0.25	0.17	**0.14**	**0.10**	**0.08**	**0.07**	**0.05**	**113**	110	61	**37**	22	**11**	**9**	**7**
OST	0.35	**0.24**	**0.16**	**0.14**	**0.10**	**0.08**	**0.07**	**0.05**	**113**	**109**	**58**	42	**17**	**11**	**9**	**7**
Jasper																
Id	18.20	12.53	7.88	5.70	3.87	2.14	1.41	1.01	1907	1220	732	559	241	160	84	55
KLT	0.91	0.53	0.43	0.34	0.26	0.20	0.15	0.12	225	151	82	57	30	15	10	7
OrthOST	0.83	**0.51**	**0.40**	0.33	**0.24**	0.19	0.15	0.12	157	**84**	**46**	**34**	23	**13**	9	7
OST	**0.79**	**0.51**	0.41	**0.32**	**0.24**	**0.18**	**0.14**	**0.11**	**156**	86	48	**34**	**22**	14	**8**	**6**

The bit-rate was computed with the VM9.

Table 4
Bit-rate (in bpppb) versus SNR for *Moffett* or PSNR for *Port de Bouc* (in dB) and versus MAD of different spectral transforms for the separable ($_1$), subband ($_2$) and mixed subband ($_3$) schemes.

Bit-rate	SNR (dB)								MAD							
	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
Moffett																
KLT_1	44.47	47.95	50.43	52.37	55.26	57.90	60.48	63.07	325	206	102	65	29	16	10	6
KLT_2	45.35	48.57	50.99	52.82	55.66	58.28	60.85	63.42	340	191	113	71	33	17	9	**5**
KLT_3	45.47	48.75	51.21	53.06	55.99	58.69	61.29	63.88	346	250	190	129	88	53	19	8
$OrthOST_1$	45.53	48.77	51.16	52.95	55.75	58.35	60.91	63.48	**147**	81	61	47	**25**	**11**	**7**	6
$OrthICA_2$	45.74	48.92	51.30	53.06	55.86	58.47	61.03	63.60	169	**79**	**54**	49	26	16	9	**5**
$OrthICA_3$	45.83	49.04	51.44	53.25	56.17	58.89	61.55	64.18	305	185	104	59	34	17	10	6
OST_1	45.53	48.75	51.15	52.93	55.74	58.34	60.91	63.48	146	86	67	51	28	15	**7**	**5**
$GCGsup_2$	45.74	48.92	51.30	53.06	55.87	58.48	61.04	63.60	166	83	58	**45**	27	16	**7**	6
$GCGsup_3$	**45.88**	**49.09**	**51.47**	**53.28**	**56.21**	**58.94**	**61.62**	**64.27**	287	160	101	58	33	18	8	**5**
Port de Bouc																
KLT_1	32.93	37.00	40.04	42.54	46.59	49.93	52.88	55.61	1021	549	362	246	147	91	70	35
KLT_2	33.01	37.09	40.16	42.65	46.73	50.08	53.07	55.85	924	557	379	246	157	91	54	36
KLT_3	33.25	37.28	40.30	42.81	46.88	50.24	53.23	56.01	947	685	388	304	171	97	53	38
$OrthOST_1$	32.91	37.03	40.20	42.76	46.93	50.32	53.30	56.06	935	625	351	249	145	85	68	36
$OrthICA_2$	33.87	37.49	40.40	42.86	46.97	50.44	53.41	56.10	**679**	**463**	**308**	272	140	**77**	50	35
$OrthICA_3$	34.09	37.82	40.67	43.08	47.18	50.62	53.58	56.27	763	524	386	269	**133**	91	49	35
OST_1	33.87	37.52	40.38	42.86	47.05	50.52	53.45	56.08	746	501	312	272	145	81	52	**34**
$GCGsup_2$	33.89	37.54	40.41	42.89	47.09	50.56	53.50	56.19	758	516	331	**245**	156	86	49	36
$GCGsup_3$	**34.10**	**37.86**	**40.71**	**43.17**	**47.33**	**50.76**	**53.68**	**56.36**	859	506	373	275	160	86	**47**	**34**

The bit-rate was computed with the average first order entropy. In the mixed subband scheme, $P = 4$.

16.3 and 16.4 dB. Moreover, we can notice that the optimal transforms OrthOST and OST perform always a little better than the KLT at medium and high bit-rates: on six multispectral (resp. three hyperspectral) images the average gains of OrthOST and OST on KLT are about 0.23 and 0.28 dB (resp. 0.42 and 0.49 dB). Further, we can remark that there is an insignificant difference of performances between OrthOST and OST. This can be explained by the fact that transforms minimizing the criterion (18) must have a small value for $C_O(\mathbf{A})$, i.e., they must be close to orthogonality (see Remark 4). Therefore there is no advantage to use OST rather than the orthogonal transform OrthOST.

In examining the MAD distortion we observe that on the multispectral images tested, at medium bit-rates (i.e., between 0.25 and 1.5 bpppb), OrthOST performs worse than the KLT (see Table 1). On the other hand, on the three hyperspectral (AVIRIS) images tested, for all the

distortions measured, at medium and high bit-rates, OrthOST performs *always* better than the KLT (see Tables 2 and 3). This is a nice finding, since the optimality of OrthOST is justified only for the MSE distortion and at high bit-rates.

We have also tested the subband and the mixed subband schemes and compared them with the separable scheme. For this, as mentioned above, to estimate the bit-rate, we used the average first order entropy $(1/NP)\sum_{i=1}^{N}\sum_{\ell=1}^{P}\sum_{m=1}^{M}\pi_m H(Y_{i,\ell}^{q(m)})$, where $H(Y_{i,\ell}^{q(m)})$ is the first order entropy of the quantized coefficients $Y_{i,\ell}^{q(m)}$. The optimal (resp. orthogonal optimal) transforms of the subband and mixed subband schemes are returned by the algorithm GCGsup for *generalized coding gain supremum* (resp. OrthICA for *Orthogonal ICA*). Both algorithms GCGsup and OrthICA are described in [20]. We only consider the MSE and MAD distortions and we use three levels of decomposition for the 2-D DWT. The decimator factor P is equal to 4 in the mixed subband scheme. In our tests (see Table 4), we observed (1) an average gain of about 0.1 dB (resp. 0.85 dB) for the subband scheme on the separable scheme for both optimal transforms OrthICA and GCGsup (resp. KLT) and (2) an average gain of about 0.37 and 0.42 dB (resp. −0.1 dB) for the mixed subband scheme on the subband scheme for OrthICA and GCGsup (resp. KLT) on hyperspectral images. For multispectral images, we observed (1) an average gain of about 0.06 dB (resp. 0.08 dB) for the subband scheme on the separable scheme for both optimal transforms OrthICA and GCGsup (resp. KLT) and (2) an average gain of about 0.21 and 0.22 dB (resp. 0.15 dB) for the mixed subband scheme on the subband scheme for OrthICA and GCGsup (resp. KLT). This was expected since it is well known that after the DWT, it remains a few dependencies between adjoining wavelet coefficients [23]. In these tests the bit-rate was estimated without taking into account the size of the bit stream occupied by the inverse spectral transform matrices. But when it is counted, it significantly increases the bit-rate of hyperspectral images and for such images, there is no gains (but losses) when the subband scheme or the mixed subband scheme is used instead of the separable scheme, except for the KLT with the subband scheme. This last fact is not surprising, since the KLT is not the optimal transform of the separable scheme, even when the data are Gaussian (see [34]). For the multispectral PLEIADES images, when the size occupied by the inverse matrices is counted, it is negligible in the subband scheme and reduces of about 0.1 dB the gain in the mixed subband scheme with $P = 4$.

7. Conclusion

In this paper, we have studied the problem of finding optimal spectral transforms in coding of multi- and hyperspectral images, for a compression scheme that is compatible with the JPEG2000 Part 2 standard. We clarified the criterion that gives, when minimized, an optimal transform under high-rate entropy constraint scalar quantization hypothesis and when one scalar quantizer per subband and per component is applied. We also introduced two variants of the compression scheme that are not JPEG2000 compatible, and the associated criteria which are minimized by optimal transforms. Then we gave two new algorithms that return the spectral transforms that minimize the JPEG2000 compatible criterion, one under the constraint of orthogonality, the other without no constraint, but invertibility. Finally, we have tested the optimal transforms of the three compression schemes on satellite multi- and hyperspectral images and found that for hyperspectral images the orthogonal optimal transform performs a little better than the KLT for four distortion measures that permit to evaluate the performances of the codec in applications of hyperspectral images like classifications or target detections. Moreover, there is no advantage to use the variants of the compatible JPEG2000 compression scheme. However, the computational complexity of the optimal transform is too heavy for actual applications and we propose in other papers different approaches to reduce the computational burden.

In order to be complete, we recall in appendixes well known results that can be found e.g., in [26].

Acknowledgments

The authors are grateful to anonymous reviewers and Pierre Duhamel for their helpful comments as well as Jean-Louis Gutzwiller who took part in the simulation work.

Appendix A. Justification of the assumption $\mathscr{H}_1$

The assumption $\mathcal{H}_1$ can be justified under the following conditions. C_1: the random vector $\mathbf{Y} = (Y_1, \ldots, Y_N)^T$ has a continuous probability density function (pdf) $f_{\mathbf{Y}}$; C_2: separable high-rate quantization is achieved, meaning that the quantization steps $\mathbf{h} = (h_i)_{1\le i\le N}$ of the N components are small with respect to the variations of $f_{\mathbf{Y}}$ (i.e. $f_{\mathbf{Y}}(\mathbf{y}+\mathbf{h}) \simeq f_{\mathbf{Y}}(\mathbf{y})$, $\forall \mathbf{y} \in \mathbb{R}^N$); C_3 (*centroid condition*): for any cell S of the separable N-D quantizer, the dequantized value $\mathbf{Y}^q$ associated with S satisfies $\mathbf{Y}^q = E[\mathbf{Y}|\mathbf{Y} \in S]$. Indeed, if both the two conditions C_1 and C_2 hold, then the pdf $f_{\mathbf{Y}}$ can be considered as quasi-constant in the hypercube $\mathbf{Y}^q + \prod_{i=1}^{N}[-h_i/2, h_i/2]$. Further, if the condition C_3 holds, then the conditional law of the quantization noise $\mathbf{b} = \mathbf{Y} - \mathbf{Y}^q$ knowing the dequantized value $\mathbf{Y}^q$ satisfies $f_{\mathbf{b}|\mathbf{Y}^q}(\mathbf{u}) \simeq 1/\prod_{i=1}^{N} h_i$ if $\mathbf{u} \in \prod_{i=1}^{N}[-h_i/2\,, h_i/2]$, 0 otherwise. We see that the conditional pdf $f_{\mathbf{b}|\mathbf{Y}^q}$ does not depend on the quantized value $\mathbf{Y}^q$, hence it is equal to $f_{\mathbf{b}}$, the pdf of $\mathbf{b}$. Further the components of $\mathbf{b}$ are zero mean and (quasi) independent since their joint density is approximatively equal to the product of their marginal densities. □

Appendix B. Computation of the optimal bits allocation

Lemma B.1. *The asymptotical optimal bit allocation of the mixed subband scheme for a given maximal end-to-end distortion D_t is achieved when the distortion in each*

block of each subband per component is taken as $D^{(m)}_{j,\ell,opt} = D_t/\omega_m w^{(m)}_{j,\ell}$.

Proof. The optimal bits allocation problem for the mixed subband scheme can be stated as follows: for a targeted end-to-end distortion D_t, how can the quantizer distortions $D^{(m)}_{j,\ell}$ be distributed in each block of each subband of each component in order to minimize the total bit-rate R expressed in the relation (14)? This is a problem of minimization under the constraint $(1/NP)\sum_{j=1}^{N}\sum_{\ell=1}^{P}\sum_{m=1}^{M}\pi_m\omega_m w^{(m)}_{j,\ell}D^{(m)}_{j,\ell} \le D_t$. As it is well known, it can be transformed into a unconstrained minimization problem, by introducing a Lagrange multiplier μ. The solution of the constrained problem is then the same as the one that minimizes

$$\mathcal{L} = \frac{1}{NP}\sum_{j=1}^{N}\sum_{m=1}^{M}\sum_{\ell=1}^{P}\pi_m\left[H(Y^{(m)}_{j,\ell}) - \frac{1}{2}\log_2(cD^{(m)}_{j,\ell})\right] + \mu\left[\frac{1}{NP}\sum_{m=1}^{M}\sum_{j=1}^{N}\sum_{\ell=1}^{P}\pi_m\omega_m w^{(m)}_{j,\ell}D^{(m)}_{j,\ell} - D_t\right]$$

with respect to the distortions $D^{(m)}_{j,\ell}$ and the Lagrangian multiplier μ. In solving the system of equations obtained by canceling out the partial derivatives of this function with respect to $D^{(m)}_{j,\ell}$ and

$$\mu : \frac{\partial\mathcal{L}}{\partial D^{(m)}_{j,\ell}} = -\frac{1}{NP}\pi_m\frac{1}{2\log(2)D^{(m)}_{j,\ell}} + \mu\frac{1}{NP}\pi_m\omega_m w^{(m)}_{j,\ell},$$

$$\frac{\partial\mathcal{L}}{\partial\mu} = \frac{1}{NP}\sum_{m=1}^{M}\sum_{j=1}^{N}\sum_{\ell=1}^{P}\pi_m\omega_m w^{(m)}_{j,\ell}D^{(m)}_{j,\ell} - D_t,$$

we find that the optimal bits allocation is achieved when the distortions satisfy $D^{(m)}_{j,\ell,opt} = D_t/\omega_m w^{(m)}_{j,\ell}$. □

Appendix C. Some tested images

Some images used in our tests are shown in Figs. 2 and 3.

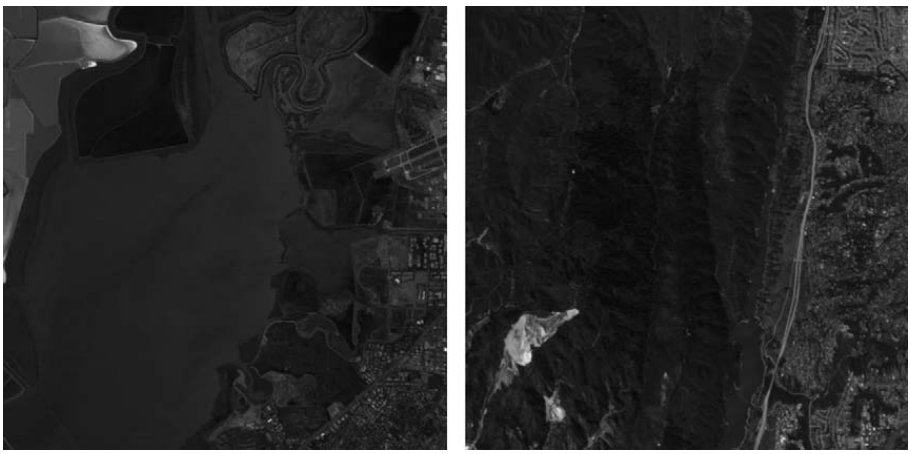


Fig. 2. From left to right: *Moffett*, *Jasper*.

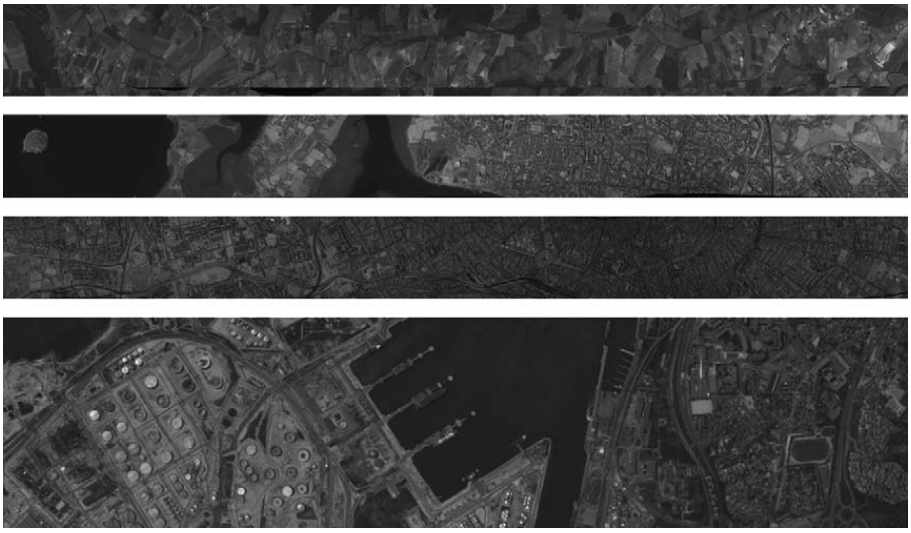


Fig. 3. From up to down *Moissac*, *Vannes*, *Toulouse*, *Port-de-Bouc*.

References

[1] P.L. Dragotti, G. Poggi, A.R.P. Ragozini, Compression of multispectral images by three-dimensional SPIHT algorithm, IEEE Transactions on Geoscience and Remote Sensing 38 (1) (2000) 416–428.

[2] A. Said, W.A. Pearlman, A new fast and efficient image codec based on set partitioning in Hierarchical trees, IEEE Transactions on Circuits and Systems for Video Technology 6 (3) (1996) 243–250.

[3] J. Vaisey, M. Barlaud, M. Antonini, Multispectral image coding using lattice VQ and the wavelet transform, in: Proceedings of the IEEE International Conference on Image Processing, Chicago (USA), October 1998, pp. 307–311.

[4] M. Antonini, M. Barlaud, P. Mathieu, I. Daubechies, Image coding using wavelet transform, IEEE Transaction on Image Processing 1 (2) (April 1992) 205–220.

[5] J.T. Rucker, J.E. Fowler, N.H. Younan, JPEG2000 coding strategies for hyperspectral data, in: Proceedings of the International Geoscience and Remote Sensing Symposium, vol. 1, July 2005, pp. 128–131.

[6] D.S. Taubman, M.W. Marcellin, JPEG2000: Image Compression Fundamentals, Standards and Practice, Kluwer Academic Publishers, Dordrecht, 2002.

[7] D.S. Taubman, High performance scalable compression with EBCOT, IEEE Transactions on Image Processing 9 (7) (July 2000) 1158–1170.

[8] E. Christophe, C. Mailhes, P. Duhamel, Best anisotropic 3-D wavelet decomposition in a rate-distortion sense, in: Proceedings of the IEEE ICASSP'06, vol. 2, May 2006, pp. II-17–20.

[9] M. Barret, I.P. Akam Bita, J.-L. Gutzwiller, F. Dalla Vedova, Lossy hyperspectral images coding with exogenous quasi optimal transforms, in: Proceedings of the Data Compression Conference, March 2009, pp. 411–419.

[10] J.E. Fowler, J.T. Rucker, Hyperspectral data exploitation: theory and applications, in: C.-I. Chang (Ed.), 3D Wavelet-based Compression of Hyperspectral ImageryWiley, Hoboken, 2007 (Chapter 14).

[11] Q. Du, J.E. Fowler, Hyperspectral image compression using JPEG2000 and principal component analysis, IEEE Geoscience and Remote Sensing Letters 4 (April 2007) 201–205.

[12] B. Penna, T. Tillo, E. Magli, G. Olmo, Transform coding techniques for lossy hyperspectral data compression, IEEE Transactions on Geoscience and Remote Sensing 45 (5) (May 2007).

[13] Q. Du, J.E. Fowler, Low-complexity principal component analysis for hyperspectral image compression, International Journal of High Performance Computing Applications, 22 (November 2008) 438–448.

[14] C. Thiebaut, D. Lebedeff, C. Latry, Y. Bobichon, On-board compression algorithm for satellite multispectral images, in: Proceedings of the Data Compression Conference, Snowbird, March 28–30, 2006.

[15] J.-Y. Huang, P.M. Schultheiss, Block quantization of correlated Gaussian random variables, IEEE Transactions on Communication COM-11 (September 1963) 289–296.

[16] V.K. Goyal, J. Zhuang, M. Vetterli, Transform coding with backward adaptive updates, IEEE Transactions on Information Theory 46 (July 2000) 1623–1633.

[17] V.K. Goyal, Theoretical foundations of transform coding, IEEE Signal Processing Magazine 18 (5) (September 2001) 9–21.

[18] M. Effros, H. Feng, K. Zeger, Suboptimality of the Karhunen–Loève transform for transform coding, IEEE Transactions on Information Theory 50 (8) (August 2004) 1605–1619.

[19] S. Mallat, F. Falzon, Analysis of low bit rate image transform coding, IEEE Transactions on Signal Processing 46 (4) (April 1998) 1027–1042.

[20] M. Narozny, M. Barret, D.-T. Pham, ICA based algorithms for computing optimal 1-D linear block transforms in variable high-rate source coding, Signal Processing 88 (2) (February 2008) 268–283.

[21] I.P. Akam Bita, M. Barret, D.-T. Pham, Compression of multi-component satellite images using independent component analysis, Proceedings of the 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA), Lecture Notes in Computer Science, vol. 3889, Springer, Berlin, March 2006, pp. 335–342.

[22] I.P. Akam Bita, M. Barret, D.-T. Pham, Transformations linéaires optimales á hauts débits pour la compression d'images multi-composantes selon la norme JPEG2000, in: Proceedings of the 21st GRETSI Colloquium, September 2007, pp. 489–492.

[23] J. Liu, P. Moulin, Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients, IEEE Transactions on Image Processing 10 (11) (2001) 1647–1658.

[24] D.-T. Pham, Fast algorithms for mutual information based independent component analysis, IEEE Transactions on Signal Processing 52 (10) (2004) 2690–2700.

[25] E. Christophe, D. Léger, C. Mailhes, Quality criteria benchmark for hyperspectral imagery, IEEE Transactions Geoscience and Remote Sensing 43 (9) (September 2005) 2103–2114.

[26] A. Gersho, R.M. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publisher, Dordrecht, 1992.

[27] J.W. Woods, T. Naven, A filter based bit allocation scheme for subband compression of HDTV, IEEE Transactions on Image Processing 1 (3) (July 1992) 436–440.

[28] B. Usevitch, Optimal bit allocation for biorthogonal wavelet coding, in: Proceedings of Data Compression Conference, April 1996, pp. 387–395.

[29] R.M. Gray, D.L. Neuhoff, Quantization, IEEE Transactions on Information Theory 44 (6) (October 1998) 2325–2384.

[30] M. Narozny, M. Barret, D.-T. Pham, I.P. Akam Bita, Modified ICA algorithm for finding optimal transforms in transform coding, in: Proceedings of the IEEE 4th International Symposium on Image and Signal Processing Analysis, September 2005, pp. 111–116.

[31] D.-T. Pham, Entropy of random variable slightly contaminated with another, IEEE Signal Processing Letters 12 (7) (2005) 536–539.

[32] JPEG2000 Verification Model 9.1 (Technical description), ISO/IEC JTC 1/SC 29/WG 1 WG1 N2165, June 2001.

[33] Y. Shoham, A. Gersho, Efficient bit allocation for an arbitrary set of quantizers, IEEE Transactions on Acoustics Speech and Signal Processing 36 (9) (1988) 1445–1453.

[34] I.P. Akam Bita, M. Barret, D.-T. Pham, On optimal orthogonal transforms at high bit-rates using only second order statistics in multicomponent image coding with JPEG2000, Signal Processing (2009), in press, doi:10.1016/j.sigpro.2009.08.008.

Signal Processing 88 (2008) 2790–2803

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Adapted generalized lifting schemes for scalable lossless image coding

Hocine Bekkouche [a,*,1], Michel Barret [a], Jacques Oksman [b]

[a] Information, Multimodality and Signal Team, SUPELEC, 2 rue É. Belin, 57070 Metz, France
[b] Signals and Electronic Systems Team, SUPELEC, Plateau de Moulon, 91192 Gif-sur-Yvette, France

ARTICLE INFO

Article history:
Received 29 October 2007
Received in revised form
2 June 2008
Accepted 4 June 2008
Available online 11 June 2008

Keywords:
Wavelets
Adapted filter banks
Adapted lifting scheme
Adaptive filtering
Lossless image compression
Still image compression
Progressive coding
Multi-resolution analysis

ABSTRACT

Still image coding occasionally uses linear predictive coding together with multi-resolution decompositions, as may be found in several papers. Those related approaches do not take into account all the information available at the decoder in the prediction stage. In this paper, we introduce an adapted generalized lifting scheme in which the predictor is built upon two filters, leading to taking advantage of all the available information. With this structure included in a multi-resolution decomposition framework, we study two kinds of adaptation based on least-squares estimation, according to different assumptions, which are either a global or a local second order stationarity of the image. The efficiency in lossless coding of these decompositions is shown on synthetic images and their performances are compared with those of well-known codecs (S + P, JPEG-LS, JPEG2000, CALIC) on actual images. Four images' families are distinguished: natural, MRI medical, satellite and textures associated with fingerprints. On natural and medical images, the performances of our codecs do not exceed those of classical codecs. Now for satellite images and textures, they present a slightly noticeable (about 0.05–0.08 bpp) coding gain compared to the others that permit a progressive coding in resolution, but with a greater coding time.

1. Introduction

The lossless image compression finds applications in satellite and medical image processing, where a lossy or near lossless coding is not satisfactory. However, in many applications of lossless coding, from time to time, lossless at full resolution is not possible because the transmission channel has a limited bandpass and then coding with a smaller resolution is better than no transmission at all. In other applications, customers need lossless coding at full resolution and other ones are satisfied with smaller resolutions of the same images. Therefore, embedded progressive coding from low resolution to lossless full resolution can be a good compromise in many applications. This coding allows to reconstruct from a truncated bit flow a decompressed image, which has a smaller resolution than the encoded one. As and when the data are received, the user is capable of enhancing the image resolution, until it reaches the original quality and resolution.

It is well known that bi-orthogonal wavelet decompositions are efficient for lossy and near lossless image compression [1], this is why they are used in the ISO JPEG2000 standard. The lifting scheme, introduced by Sweldens [2] in order to construct wavelet decompositions by a simple, reversible and fast process, found quickly its main application in lossless image compression. In this case, a nonlinear filter bank with critical sampling and perfect reconstruction is obtained, with nonlinearities

* Corresponding author.
E-mail addresses: h_bekkouche@yahoo.com, Hocine.Bekkouche@supelec.fr (H. Bekkouche), Michel.Barret@supelec.fr (M. Barret), Jacques.Oksman@supelec.fr (J. Oksman).

[1] This work was partially supported by the Lorraine Region.

0165-1684/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2008.06.003

which are limited to truncations (i.e., rounding to the nearest integer) [3]. Moreover, Daubechies and Sweldens showed that any bi-orthogonal wavelet decomposition with FIR (finite impulse response) filters can be represented by a lifting scheme [4] and, therefore, all the well-known wavelets used in lossy image codecs can be quite closely approximated by integer-to-integer wavelets. The performances in lossy and lossless image compression of integer-to-integer wavelets and the S + P transform by Said and Pearlman [5] are evaluated in [6]. Hampson and Pesquet [7] proposed a structure which is more general than the lifting scheme, with an arbitrary number of channels and arbitrary nonlinear filters. It is interesting to note the simplicity of this structure and the way the perfect reconstruction is performed in an inherent manner by a synthesis filter bank "mirror" of the analysis filter bank, as in the lifting scheme. That structure, with nonlinear prediction filters based on image segmentations, has been applied for still image and video coding by Amonou and Duhamel [8].

In the standard wavelet decompositions, the filter coefficients are fixed: they do not adapt to the image as best possible. However, the lifting scheme gives an interpretation in terms of estimation (or prediction) of perfect reconstruction filter banks, associated with multi-resolution decompositions. Now, linear prediction coding (LPC) proved its great efficiency for speech coding; it found applications in mobile telephones. Therefore it is natural to study LPC in image coding. About 15 years ago, adaptive linear predictions using least-squares estimation (LSE) algorithms were tested for image compression (see [9] and its bibliography, or later [10]), but they were not associated with dyadic decompositions and consequently they were not suitable for progressive coding. More recently, Gerek and Çetin [11] used the lifting scheme with adaptive predict steps: the filter coefficients were updated to each pixel of the image, thanks to a conventional stochastic gradient algorithm, in order to minimize the variance of the detail signal. Boulgouris et al. [12], expressed each filter of the optimal M-subband analysis filter bank as a function of the power spectral density (PSD) of the input image. They assumed the entire image is a wide sense stationary (WSS) signal. The optimum is achieved by minimizing the mean squared error of prediction for each of the $M-1$ detail signals. Two kinds of parameterized models were assumed for the PSD of the image, i.e., the adaptation is optimum only if the PSD of the image belongs to a set of two models. The filters of the update steps did not adapt to the image, they were identical with those encountered in the lifting scheme of well-known wavelets. To improve the prediction whenever the global WSS assumption is invalid, the linear predictors were enhanced by nonlinear means, namely by directional post-processing in the quincunx decimation case, and by adaptive-length post-processing in the separable (row-column) decimation case. In [13], the authors chose locally, among a finite dictionary of wavelet filters, the filter that must be applied to the current pixel depending on its proximity with an outline: the closer the pixel is to an outline, the smaller the impulse response support of the analysis filter. In [14], the authors studied the optimization of a lifting scheme (for both the predict and update steps) associated with twofold quincunx decimation. They imposed constraints to the filters in order to avoid overflow and they applied their filter banks to lossy image compression.

In each of the above mentioned papers with adapted prediction filters, we can notice that all the information available at the decoder is not taken into account in the "predict" step.[2] Indeed, after the twofold decimation, the pixels of a subband, say x_2, are predicted as a linear combination of the pixels of the other subband, say x_1, and the pixels of subband x_2 are not involved in the observation vector, whereas they could be! As is done in the classical LPC. In [15], we introduced an adapted integer-to-integer multi-resolution decomposition, based on LSE and assuming global second order stationarity of the image, which takes advantage of all the information available at the decoder, and we applied it to lossless image coding. In [16], we completed this decomposition by introducing another adaptation, which assumes only local stationarity in the image. The reason that led us to carry out this study lies in the fact that the image models are not fully appropriate for entire images, they are better justified for well-chosen parts of the images taken separately. Those parts are the textured regions that can be found in most kinds of images. Then, in [17] we compared the performances of these decompositions in lossless coding of satellite and medical MRI images with well-known codecs.

In this paper, we complete the results of the conference papers [15–17] and provide more details and full proofs. First, we present the adapted generalized lifting scheme framework, which is shared both by locally and globally adapted estimation methods—we shall call them, respectively, LAE and GAE below. In Sections 3.1 and 3.2, the GAE and LAE methods are explained in details. Their efficiency in lossless coding is shown on synthetic images (Section 4) and their performances are compared with those of well-known codecs (S + P [5,18], LOCO I [19,20], CALIC [21,22], and Jasper [23]) on actual images (Section 5). We considered four families of images (natural, medical MRI, satellite and textures with fingerprints).

In the following, $\mathbb{Z}$ denotes the set of all integers. For a matrix $\mathbf{A}$, $\mathbf{A}^{\mathrm{T}}$ denotes its transpose. Underlined lower case letters denote vectors, which are identified with the column matrix of their coordinates. The symbol E denotes the mathematical expectation.

2. Adapted generalized lifting schemes

In this section we begin by presenting a short overview of the generalized lifting scheme in the mono-dimensional (1-D) case, then we extend it to the 2-D case, clarifying the integer-to-integer variant and the adaptation of the filters. Furthermore we explain how the generalized lifting scheme can be used in a multi-resolution framework,

[2] We should say "estimation" step, since it is not a prediction problem, but an estimation problem in estimation theory; nevertheless we chose the vocabulary used in filter bank theory.

permitting a progressive coding in resolution, and we specify what kind of extension is applied to the edges of the image during the filtering. As introduction, we recall in Fig. 1 the principle of the standard lifting scheme (without adaptation) for a perfect reconstruction filter bank with two subbands, L "Predict" steps (denoted $P_1(z), P_2(z), \ldots, P_L(z)$) and L "Update" steps (denoted $U_1(z)$, $U_2(z), \ldots, U_L(z)$). The synthesis filter is easily obtained from the analysis filter by a simple mirror symmetry.

2.1. Principle of the method in the 1-D case

The general diagram is presented in Fig. 2. The polyphase components of the input signal are obtained by a polyphase decomposition (i.e., a lazy wavelet transform) [24,25]. One of the components, $x_2(n)$, is estimated with the use of two filters $A(z)$ and $B(z)$, the signal of details corresponds to the error of estimation $x_2(n) - \hat{x}_2(n)$. The approximate signal x_ℓ is just the polyphase component x_1 of the input signal x. This is a generalized version of the standard lifting scheme (see Fig. 1), since the filter $B(z)$ is added in order to improve the prediction. The polyphase matrix of the analysis filter bank is equal to

$$\mathbf{R}(z) = \begin{bmatrix} 1 & 0 \\ -A(z) & 1 - B(z) \end{bmatrix} \tag{1}$$

and then, for perfect reconstruction, the polyphase matrix of the synthesis filter bank is equal to

$$\mathbf{R}^{-1}(z) = \frac{1}{1 - B(z)} \begin{bmatrix} 1 - B(z) & 0 \\ A(z) & 1 \end{bmatrix}. \tag{2}$$

In order to reconstruct the input signal perfectly with the synthesis filter bank, it is clear that the filter $B(z)$ must be causal, since only the passed samples of $x_2(n)$ are available at the decoder for the reconstruction of $\hat{x}_2(n)$. Nevertheless, the causality of the filter $A(z)$ is not required, since all the samples of the polyphase component x_1 are available as inputs of the synthesis filter bank. Another major condition for perfect reconstruction is the BIBO (bounded-input–bounded-output) stability of the filter $1/(1 - B(z))$. Indeed, small perturbations, like round-off errors, would lead to a divergence of the reconstruction algorithm when this filter is not stable. Therefore a stability test for 1-D filter [26] is required in the case of linear filter banks.

The adaptation of the filters and the properties of the generalized lifting scheme are very similar in the 1-D and 2-D cases, so in order to be brief, we present them only in the 2-D case.

2.2. Generalized lifting scheme for 2-D signals

The extension to the 2-D case of the previous generalized lifting scheme can be carried out in two different

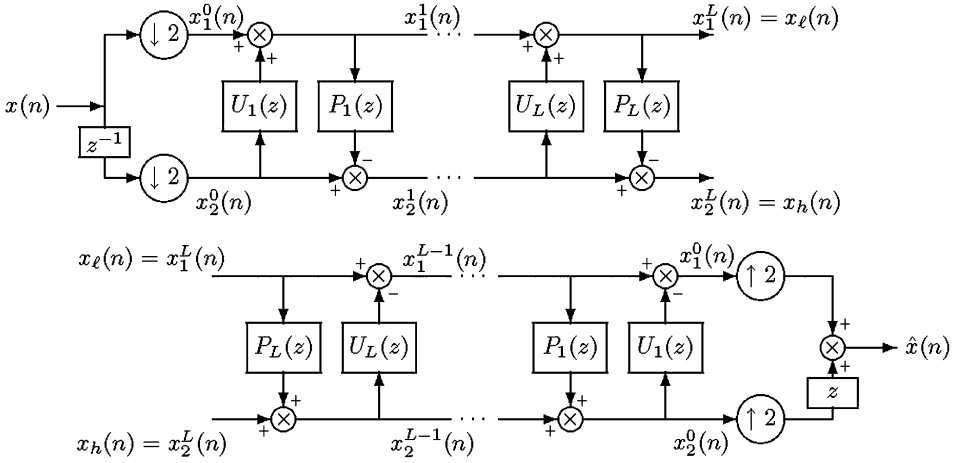


Fig. 1. Standard lifting scheme for perfect reconstruction filter bank with two subbands. Top: analysis filter bank; bottom: associated synthesis filter bank.

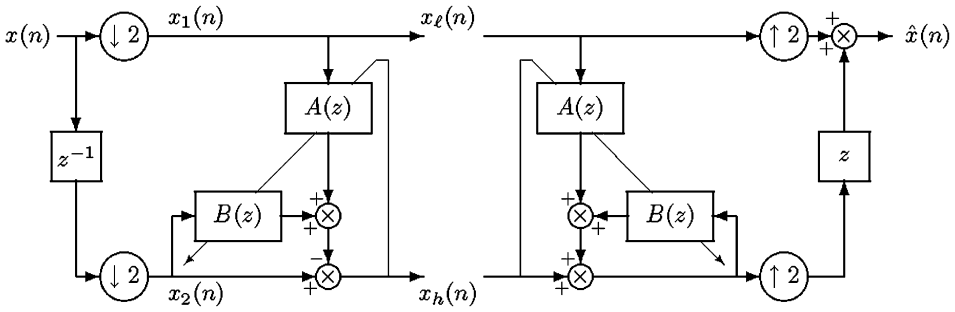


Fig. 2. Predict step of the generalized lifting scheme. Compared to the standard lifting scheme, a second filter $B(z)$ is introduced. The filters A and B adapt to the input data.

ways, according to the polyphase decomposition which can be either a twofold quincunx decimation or a twofold separable (row/column) decimation. In both cases, both of the 2-D filters $A(z_1,z_2)$ and $B(z_1,z_2)$ are not separable *a priori*. As in the 1-D case, the causality or semi-causality of the filter $B(z_1,z_2)$ is required for perfect reconstruction. Moreover, the 2-D stability test [26] of the filter $1/(1-B(z_1,z_2))$ is also required in the linear case. We shall see in Section 2.4 (Remark 2) that the stability test is *not* required for the *integer-to-integer* generalized lifting scheme.

The filters in Fig. 3 are expressed in terms of their transfer functions

$$A(z_1,z_2)=\sum_{(i,j)\in\Delta_1} a_{ij}z_1^{-i}z_2^{-j}$$

and

$$B(z_1,z_2)=\sum_{(i,j)\in\Delta_2} b_{ij}z_1^{-i}z_2^{-j}, \tag{3}$$

where Δ_1 is any subset of $\mathbb{Z}^2$ (no condition of causality is imposed) and Δ_2 is a subset of a non-symmetrical half plane (NSHP) of $\mathbb{Z}^2$ to ensure the semi-causality of $B(z_1,z_2)$.

The prediction of $x_2(m,n)$ is given by the equation

$$\hat{x}_2(m,n)=\sum_{(i,j)\in\Delta_1} a_{ij}x_1(m-i,n-j)+\sum_{(i,j)\in\Delta_2} b_{ij}x_2(m-i,n-j) \tag{4}$$

and the criterion used for the filter coefficients adaptation is the minimization of the mean squared error $\mathrm{E}[|x_2(m,n)-\hat{x}_2(m,n)|^2]$.

2.3. Adaptation of the filters

To solve the problem of 2-D LSE (4), we first scan the elements of subsets Δ_1 and Δ_2 in a preset order. We build thus, from the respective families of pixels $x_1(m-i,n-j)|_{(i,j)\in\Delta_1}$ and $x_2(m-i,n-j)|_{(i,j)\in\Delta_2}$, observation vectors $\underline{y}_1(m,n)$ and $\underline{y}_2(m,n)$ of respective sizes r_1 and r_2. The scans chosen for the description of Δ_1 and Δ_2 impose an organization of the coefficients $a_{i,j}|_{(i,j)\in\Delta_1}$ and $b_{ij}|_{(i,j)\in\Delta_2}$ of filters $A(z_1,z_2)$ and $B(z_1,z_2)$ as elements of the respective vectors $\underline{a}$ and $\underline{b}$ so that

$$\hat{x}_2(m,n)=[\underline{a}^{\mathrm{T}},\underline{b}^{\mathrm{T}}]\begin{bmatrix}\underline{y}_1(m,n)\\ \underline{y}_2(m,n)\end{bmatrix}=\underline{c}^{\mathrm{T}}\underline{y}(m,n) \tag{5}$$

with $\underline{c}^{\mathrm{T}}=[\underline{a}^{\mathrm{T}},\underline{b}^{\mathrm{T}}]$ and $\underline{y}(m,n)^{\mathrm{T}}=[\underline{y}_1(m,n)^{\mathrm{T}},\underline{y}_2(m,n)^{\mathrm{T}}]$.

While applying, for example the principle of orthogonality, one finds that vector $\underline{c}$, of dimension $r=r_1+r_2$, is the solution of the Yule–Walker equations

$$\Gamma_Y\,\underline{c}=\underline{\gamma}_{yx} \tag{6}$$

with $\Gamma_Y=\mathrm{E}[\underline{y}(m,n)\,\underline{y}(m,n)^{\mathrm{T}}]$ and $\underline{\gamma}_{yx}=\mathrm{E}[\underline{y}(m,n)x_2(m,n)]$. We shall see thereafter how to estimate these mathematical expectations from the data. We shall distinguish two cases, according to the assumption of WSS made on the signal $x(m,n)$: global (i.e., on the entire image) in Section 3.1 or local in Section 3.2.

Remark 1. For a WSS signal $x(m,n)$, it is well known in estimation theory that the residue of the linear LSE, which corresponds to the image of details

$$x_h(m,n)=x_2(m,n)-\hat{x}_2(m,n), \tag{7}$$

is uncorrelated with the observations $x_1(m-i,n-j)|_{(i,j)\in\Delta_1}$ and $x_2(m-i,n-j)|_{(i,j)\in\Delta_2}$. Moreover, by extending well-known results on linear LSE to the 2-D case, it is clear that if signal $x(m,n)$ is WSS and if the supports Δ_1 and Δ_2 tend toward infinite ones (i.e., $\Delta_1=\mathbb{Z}^2$ and Δ_2 is a NSHP), then the residue $x_h(m,n)$ tends toward a white noise. That explains why, when the supports are large enough, the coefficients close to $x_h(m,n)$ are very slightly correlated between them and with the coefficients of subband x_ℓ. This fact is valid even when the assumption of second order stationarity is just locally satisfied.

2.4. Integer-to-integer generalized lifting scheme

When implemented with a fixed (or a floating) point arithmetic processor, the reversibility of the decomposition is not ensured with Eq. (7), since the filter coefficients are not integers. To avoid this problem, it is enough to round the estimation to the nearest integer before removing it from the exact value in order to generate the detail signal, which becomes

$$x_h(m,n)=x_2(m,n)-\left\lfloor\sum_{(i,j)\in\Delta_1} a_{i,j}x_1(m-i,n-j)+\sum_{(i,j)\in\Delta_2} b_{i,j}x_2(m-i,n-j)+\frac{1}{2}\right\rfloor, \tag{8}$$

where $\lfloor x\rfloor$ denotes the largest integer not greater than x. Thus, we obtain the integer-to-integer generalized lifting scheme.

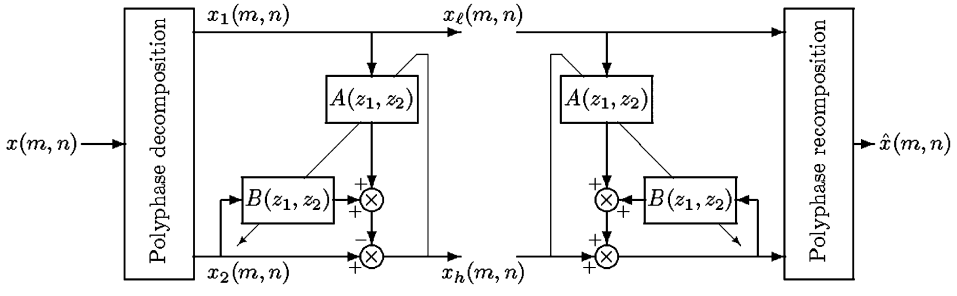


Fig. 3. Predict step of the generalized lifting scheme in the 2-D case. The additional filter $B(z_1,z_2)$ is introduced in comparison with the standard lifting scheme. Each of the filters A and B are adapted to the input signal.

Remark 2. It is important to notice here that if the coder and the decoder use the same floating point arithmetic (with the same round off rule) to calculate the expression between $\lfloor$ and $\rfloor$ in Eq. (8), then the decomposition is perfectly reversible, since $x_h(m,n)$ and $x_2(m,n)$ are integers and arithmetic operations between integers are perfectly reversible (when no overflow is detected) on any processor. Moreover, this reasoning shows that the perfect reversibility remains valid even when filter $1/(1-B(z_1,z_2))$ is unstable. Therefore the BIBO stability test is not required, for the integer-to-integer generalized lifting scheme to ensure perfect reconstruction.

2.5. Multi-resolution decomposition

The multi-resolution decomposition of a 2-D signal is based on the decomposition shown in Fig. 3. We first explain one level of decomposition in four subbands: the input signal $x(m,n)$ is divided into two subband signals x_1 and x_2 by a polyphase decomposition; the signal of approximation x_ℓ is then equal to x_1 and the signal of details x_h is obtained according to relation (7) (or (8) for integer-to-integer decomposition). The same process is applied to x_ℓ in order to generate the two signals $x_{\ell\ell}$ and $x_{\ell h}$ of subbands LL and LH (of course, if the twofold decimation is separable then the polyphase decomposition is alternatively applied to rows and columns). However, contrary to what is usually done—in particular in the case of dyadic wavelet decomposition—to generate the two signals $x_{h\ell}$ and x_{hh} of subbands HL and HH, only a polyphase decomposition is applied to signal x_h, as indicated in Fig. 4. Indeed, after obtaining the signal x_h according to the process in Fig. 3, the correlation between adjoining samples of x_h almost vanishes (see Remark 1). Consequently, the reduction in variance obtained by applying a predict step after the polyphase decomposition to generate x_{hh} is so weak that we preferred not to use it, thus reducing the complexity of the codec. In order to have a decomposition on several levels, the same process is applied recursively on the signal of approximation $x_{\ell\ell}$.

2.6. Optional update and predict steps with fixed coefficients filters

In order to allow a progressive coding in resolution, it is necessary to avoid the artifacts (spectral aliasing) due to the twofold decimation when the resolution becomes low. For this we recommend to apply some update and predict steps with fixed coefficient filters corresponding to well-known wavelet decomposition in the adapted generalized lifting scheme, between the polyphase decomposition and the adapted predict step. In our experiments, we chose the *S*-transformation, because of its low complexity; it consists in the two steps (for mono-dimensional signal $x(n)$): $x_h(n) = x(2n) - x(2n+1)$ and $x_\ell(n) = x(2n+1) + \lfloor x_h(n)/2 \rfloor$.

2.7. Symmetrical or zero-padding extension of the image

An image is a 2-D signal with finite support, therefore a strategy must be adopted to filter near the edges of the image. Generally, the strategy consists in either zero padding, or a periodical extension, or a symmetrical extension. In our tests, we chose the last one when the decimation is separable, since it avoids discontinuities and thus generally gives better performances in coding. Nevertheless, with quincunx decimation, we adopted zero padding to simplify the implementation of the codec.

3. Two approaches of least-squares estimation

We distinguish two kinds of adaptation for the filters $A(z_1,z_2)$ and $B(z_1,z_2)$ according to the assumptions of global or local WSS of the encoded image.

3.1. Globally adapted estimation (GAE) method

3.1.1. Choice of the criterion

The input signal is supposed to be a realization of a 2-D WSS stochastic process. The subset $\Delta_1 \subset \mathbb{Z}^2$ is a rectangle

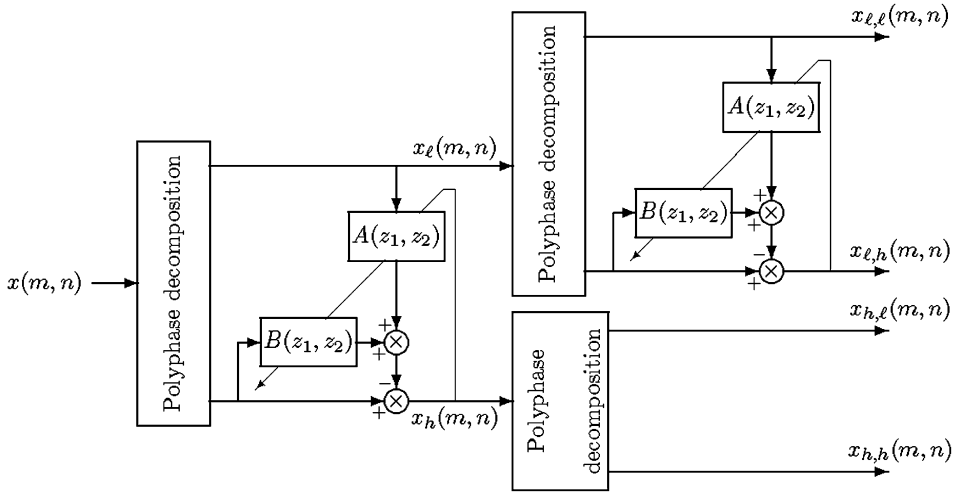


Fig. 4. One level of decomposition of the image with the adapted generalized lifting scheme.

centered in $(0,0)$: $\Delta_1 = \Delta_1'$ with

$$\Delta_1' = \{(i,j) \in \mathbb{Z}^2 : |i| \leqslant p \text{ and } |j| \leqslant q\} \tag{9}$$

and the subset $\Delta_2 \subset \mathbb{Z}^2$ is a bounded NSHP: $\Delta_2 = \Delta_2'$ with

$$\Delta_2' = \left\{ (i,j) \in \mathbb{Z}^2 : \begin{array}{c} (i = 0 \text{ and } 1 \leqslant j \leqslant q) \\ \text{or} \\ (1 \leqslant i \leqslant p \text{ and } |j| \leqslant q) \end{array} \right\}. \tag{10}$$

The vectors, $\underline{y}_1(m,n) = x_1(m+p,n+q), \ldots, x_1(m+p, n-q), x_1(m+p-1,n+q), \ldots, x_1(m+p-1,n-q), \ldots, x_1(m-p+1,n+q), \ldots, x_1(m-p+1,n-q), x_1(m-p,n+q), \ldots, x_1(m-p,n-q)]^{\mathrm{T}}$ of dimension $r_1 = (2p+1)(2q+1)$ and $\underline{y}_2(m,n) = [x_2(m,n-1), \ldots, x_2(m,n-q), x_2(m-1, n+q), \ldots, x_2(m-1,n-q), \ldots, x_2(m-p+1,n+q), \ldots, x_2(m-p+1,n-q), x_2(m-p,n+q), \ldots, x_2(m-p,n-q)]^{\mathrm{T}}$ of dimension $r_2 = p(2q+1)+q$, contain the values of the respective samples $x_1(m-i,n-j)|_{(i,j)\in\Delta_1}$ and $x_2(m-i, n-j)|_{(i,j)\in\Delta_2}$. The vectors $\underline{a}$ and $\underline{b}$ can be expressed as $\underline{a} = [a_{-p,-q}, \ldots, a_{-p,q}, a_{-p+1,-q}, \ldots, a_{-p+1,q}, \ldots, a_{p,-q}, \ldots, a_{p,q}]^{\mathrm{T}}$ and $\underline{b} = [b_{0,1}, \ldots, b_{0,q}, b_{1,-q}, \ldots, b_{1,q}, \ldots, b_{p,-q}, \ldots, b_{p,q}]^{\mathrm{T}}$.

After the polyphase decomposition applied to the image $x(m,n)$ of dimension $M \times N$, the images of the subbands $x_1(m,n)$ and $x_2(m,n)$ have the respective dimensions $M_1 \times N_1$ and $M_2 \times N_2$, the support of $x_2(m,n)$ is then $0 \leqslant m < M_2$ and $0 \leqslant n < N_2$. According to the orders of the filters $A(z_1,z_2)$ and $B(z_1,z_2)$, the vectors $\underline{y}_1(m,n)$ and $\underline{y}_2(m,n)$ can have supports which are located more or less beside the horizon of observation of x, hence in order to compute the actual matrices Γ_Y and $\underline{\gamma}_{yx}$ of relation (6), the observations must be extended beside the horizon of observation. Three methods have been tested.

Pre-windowed method: The subband signals $x_1(m,n)$ and $x_2(m,n)$ are supposed to be null for $m<0$ or $n<0$ and no hypothesis is made on their values for $m \geqslant M_2$ or $n \geqslant N_2$.

Autocorrelation method: The subband signals $x_1(m,n)$ and $x_2(m,n)$ are supposed to be null beside the horizon of observation, i.e., for $m<0$ or $n<0$ and for $m \geqslant M_2$ or $n \geqslant N_2$.

Covariance method: No hypothesis is made on the values of x_1 and x_2 beside the horizon of observation.

Our simulations on actual images show that the autocorrelation method gives quite often and on average first order entropies slightly greater than each of the two other methods. In the continuation of our simulations, we use the pre-windowed method, which gives often and on average performances similar to the covariance method with a weaker complexity.[3]

The criterion to be minimized is thus given by

$$J_1 = \frac{1}{(M_2-p)(N_2-q)} \sum_{m=0}^{M_2-p-1} \sum_{n=0}^{N_2-q-1} x_h(m,n)^2 \simeq \mathrm{E}[x_h(m,n)^2]. \tag{11}$$

Let us gather the elements $x_h(m,n)|_{0\leqslant m<M_2-p,\, 0\leqslant n<N_2-q}$ in the vector $\underline{\mathscr{X}}_h$ by scanning x_h row after row: $\underline{\mathscr{X}}_h = [x_h(0,0)\cdots x_h(0,N_2-q-1)\cdots x_h(M_2-p-1,0)\cdots x_h(M_2-p-1,N_2-q-1)]^{\mathrm{T}}$ we then get the relation $\underline{\mathscr{X}}_h = \underline{\mathscr{X}}_2 - \mathscr{Y}\,\underline{c}$, with $\underline{\mathscr{X}}_2 = [x_2(0,0)\cdots x_2(0,N_2-q-1)\cdots x_2(M_2-p-1,0)\cdots x_2(M_2-p-1,N_2-q-1)]^{\mathrm{T}}$ and the matrix of dimension $(M_2-p)(N_2-q)\times r$, which is equal to $\mathscr{Y} = [\underline{y}(0,0)\cdots\underline{y}(0,N_2-q-1)\cdots\underline{y}(M_2-p-1,0)\cdots\underline{y}(M_2-p-1,N_2-q-1)]^{\mathrm{T}}$. Criterion (11) to be minimized can be written $(M_2-p)(N_2-q)J_1 = \|\underline{\mathscr{X}}_h\|_2^2 = \|\underline{\mathscr{X}}_2 - \mathscr{Y}\,\underline{c}\,\|_2^2 = [\underline{\mathscr{X}}_2 - \mathscr{Y}\,\underline{c}\,]^{\mathrm{T}}[\underline{\mathscr{X}}_2 - \mathscr{Y}\,\underline{c}]$ and the vector $\underline{\hat{c}}$ is the solution of the Yule–Walker equations

$$\mathscr{Y}^{\mathrm{T}}\mathscr{Y}\,\underline{\hat{c}} = \mathscr{Y}^{\mathrm{T}}\underline{\mathscr{X}}_2. \tag{12}$$

3.1.2. Complexity of the method

In the GAE method, the stage that requires most of the arithmetic operations is the calculation of the matrix $\mathscr{Y}^{\mathrm{T}}\mathscr{Y}$ of the Yule–Walker equations. Therefore, it is important to reduce its complexity by using the great redundancy that exists between its elements. The fast computation of covariance matrix associated to 2-D signals given in [10] does not apply here. So, we present in Appendix A the details of the proposed method, it is an extension to the 2-D case of the pre-windowed method given in [27] and we show that the computation of the matrix $\mathscr{Y}\mathscr{Y}^{\mathrm{T}}$ costs $2[2(4q+1)(3p+1)+1]M_2N_2 + o(M_2N_2)$ arithmetic operations.

3.2. LAE method

3.2.1. Choice of the criterion

The input signal is supposed to be a realization of a 2-D stochastic process, whose statistics of orders 1 and 2 are locally stationary. We apply then the recursive least squares (RLS) algorithm. For this, we introduce a forgetting factor α, $0<\alpha\ll 1$, and the criterion to be minimized (i.e., the local estimation of the variance of the prediction error) is equal to

$$J_2(m,n) = \sum_{i=0}^{m-1}\sum_{j=0}^{N_2-1} \alpha^{N_2(m-i)+n-j} x_h(i,j)^2 + \sum_{j=0}^{n} \alpha^{n-j} x_h(m,j)^2 \tag{13}$$

with

$$x_h(i,j) = x_2(i,j) - \underline{c}^{\mathrm{T}}(m,n)\,\underline{y}(i,j). \tag{14}$$

We can notice in this relation that we associated the index (m,n) with the vector $\underline{c}$ and the index (i,j) with the other variables. It is due to the nature of the optimization: for each (m,n), $\underline{c}(m,n)$ depends on $\underline{y}(i,j)$ and $x_2(i,j)$ for $0\leqslant i<m$ and $0\leqslant j<N_2$ and also for $i=m$ and $0\leqslant j\leqslant n$. The forgetting factor is applied in a way specific to the method of scanning the image (from left to right and from top to bottom).

For the LAE method, the supports Δ_1 and Δ_2 of the filters used in relations (4) and (8) are subsets of Δ_1' and Δ_2' given, respectively, in relations (9) and (10). Indeed, in an image the pixels which contain the most useful information for the estimation of the current pixel are generally its closest neighbors. It is thus natural to reorder the components of vectors $\underline{y}_1(m,n)$ and $\underline{y}_2(m,n)$, so that

[3] We emphasize that the three methods of extension at the edges of the image above mentioned are useful only for clarifying the matrices Γ_Y and $\underline{\gamma}_{yx}$ in relation (6). For instance, with the pre-windowed method, the Yule–Walker equations (6) become Eq. (12). Now, when the optimal filter is applied according to relations (4) or (8), we use the extensions at the edges of the image as indicated in Section 2.7.

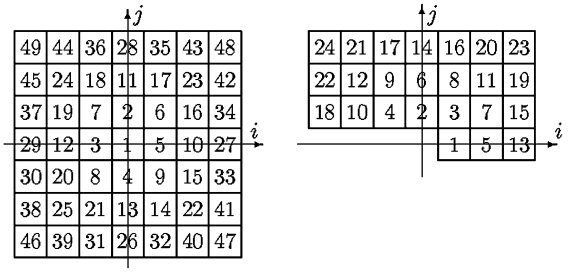


Fig. 5. Order of the samples in the observation vector for $p \leqslant 3$; left: in Δ_1'; right: in Δ_2' (LAE method).

their new order is an increasing function of the distance to the current pixel. Therefore, only r_1 coefficients of Δ_1' are retained, those numbered from 1 to r_1 in the left diagram of Fig. 5, and only r_2 coefficients of Δ_2' are retained, those numbered from 1 to r_2 on the right diagram of Fig. 5. So, it is now possible to decrease the order of the filters and consequently the computing time without reducing the performances.

Using relation (14) in Eq. (13), the criterion J_2 can be expressed as

$$J_2(m,n) = \kappa(m,n) - 2\,\underline{c}(m,n)^{\mathrm{T}}\,\underline{\Theta}(m,n) + \underline{c}(m,n)^{\mathrm{T}}\boldsymbol{\Phi}(m,n)\,\underline{c}(m,n) \tag{15}$$

with $\kappa(m,n) = \sum_{i=0}^{m-1}\sum_{j=0}^{N_2-1}\alpha^{N_2(m-i)+n-j}x_2(i,j)^2 + \sum_{j=0}^{n}\alpha^{n-j}x_2(m,j)^2$ and

$$\underline{\Theta}(m,n) = \sum_{i=0}^{m-1}\sum_{j=0}^{N_2-1}\alpha^{N_2(m-i)+n-j}x_2(i,j)\,\underline{y}(i,j) + \sum_{j=0}^{n}\alpha^{n-j}x_2(m,j)\,\underline{y}(m,j), \tag{16}$$

$$\boldsymbol{\Phi}(m,n) = \sum_{i=0}^{m-1}\sum_{j=0}^{N_2-1}\alpha^{N_2(m-i)+n-j}\,\underline{y}(i,j)\,\underline{y}(i,j)^{\mathrm{T}} + \sum_{j=0}^{n}\alpha^{n-j}\,\underline{y}(m,j)\,\underline{y}(m,j)^{\mathrm{T}}. \tag{17}$$

The criterion $J_2(m,n)$ can now be minimized by cancelling its partial derivatives with respect to $\underline{c}(m,n)$: $\partial J_2(m,n)/\partial\,\underline{c}(m,n) = -2\,\underline{\Theta}(m,n) + 2\boldsymbol{\Phi}(m,n)\,\underline{c}(m,n)$. We obtain the Yule–Walker equations:

$$\boldsymbol{\Phi}(m,n)\,\underline{c}(m,n) = \underline{\Theta}(m,n) \tag{18}$$

which have the solution $\underline{\hat{c}}(m,n) = \boldsymbol{\Phi}(m,n)^{-1}\,\underline{\Theta}(m,n)$. From relations (16) and (17) we deduce the recursive expressions[4]

$$\boldsymbol{\Phi}(m,n) = \alpha\boldsymbol{\Phi}(m,n-1) + \underline{y}(m,n)\,\underline{y}(m,n)^{\mathrm{T}}, \tag{19}$$

$$\underline{\Theta}(m,n) = \alpha\,\underline{\Theta}(m,n-1) + x_2(m,n)\,\underline{y}(m,n)^{\mathrm{T}}. \tag{20}$$

The RLS algorithm expresses $\boldsymbol{\Phi}^{-1}(m,n)$ in terms of $\boldsymbol{\Phi}^{-1}(m,n-1)$ in such a way that Eq. (18) can be solved without the inversion of the matrix $\boldsymbol{\Phi}(m,n)$ at each step (m,n). For this, the matrix inversion lemma [28]: $(\mathbf{S}^{-1} + \mathbf{U}\mathbf{V}^{-1}\mathbf{U}^{\mathrm{T}})^{-1} = \mathbf{S} - \mathbf{S}\mathbf{U}[\mathbf{V} + \mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U}]^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{S}$ is applied, with $\mathbf{S}^{-1} = \alpha\boldsymbol{\Phi}(m,n-1)$, $\mathbf{U} = \underline{y}(m,n)$ and $\mathbf{V}^{-1} = 1$. That leads to

$$\boldsymbol{\Phi}^{-1}(m,n) = \alpha^{-1}\boldsymbol{\Phi}^{-1}(m,n-1) - \frac{\alpha^{-1}\boldsymbol{\Phi}^{-1}(m,n-1)\,\underline{y}(m,n)\,\underline{y}(m,n)^{\mathrm{T}}\boldsymbol{\Phi}^{-1}(m,n-1)}{\alpha + \underline{y}(m,n)^{\mathrm{T}}\boldsymbol{\Phi}^{-1}(m,n-1)\,\underline{y}(m,n)}. \tag{21}$$

This relation requires an initial value $\boldsymbol{\Phi}(1,1)$. The simplest way is to take it equal to $\delta\mathbf{I}$, where δ is a small constant. Let $\underline{g}(m,n)$ be the vector

$$\underline{g}(m,n) = \frac{\boldsymbol{\Phi}^{-1}(m,n-1)\,\underline{y}(m,n)}{\alpha + \underline{y}(m,n)^{\mathrm{T}}\boldsymbol{\Phi}^{-1}(m,n-1)\,\underline{y}(m,n)}. \tag{22}$$

Eq. (21) becomes $\boldsymbol{\Phi}^{-1}(m,n) = \alpha^{-1}\boldsymbol{\Phi}^{-1}(m,n-1) - \alpha^{-1}\,\underline{g}(m,n)\,\underline{y}(m,n)^{\mathrm{T}}\boldsymbol{\Phi}^{-1}(m,n-1)$ and Eq. (22) can be changed into $\underline{g}(m,n) = \boldsymbol{\Phi}^{-1}(m,n)\,\underline{y}(m,n)$. This vector is called the *adaptation gain*, since it appears as a gain applied to the error of prediction in the equation that updates $\underline{\hat{c}}(m,n)$: $\underline{\hat{c}}(m,n) = \underline{\hat{c}}(m,n-1) + \underline{g}(m,n)[x_2(m,n) - \underline{\hat{c}}^{\mathrm{T}}(m,n-1)\,\underline{y}(m,n)]$. The procedure of adaptive decomposition is summarized in Table 1.

3.2.2. Complexity of the method

The complexity[5] of the LAE method is $O(M_2N_2(r_1+r_2)^2)$ [28] for one level of decomposition. Let us notice that this procedure of decomposition uses the 1-D RLS algorithm [28]. Nevertheless, the construction of the observation vector $\underline{y}(m,n)$, where the bi-dimensional neighborhood of sample $x_2(m,n)$ appears, gives a 2-D characteristic to this algorithm, i.e., why we kept the indexes (m,n) in the notation of the vectors.

3.3. Comparison of the two LSE methods

The major difference between the two methods is that the filter coefficients must be transmitted to the decoder with the GAE, but not with the LAE. Indeed, with the GAE method, the optimal filter is computed only during the coding, whereas with the LAE method, it is estimated for each pixel during both the coding and the decoding. This is why we obtain such a difference between the mean decoding times of the two methods (see Table 6). Another difference is that the computation of the optimal filter requires an extension at the edges of the image only with the GAE. The major similarity between GAE and LAE is that they minimize each the variance of the residue of estimation. It is the estimation of this variance that differs according to the stationarity assumptions made on the image: global or local. However, it is well known that the minimization of the variance is optimal for coding Gaussian data and is not optimal generally when the data are not Gaussian. Another important similarity is that included in a multi-resolution framework, each method permits progressive coding in resolution. Let us notice that since the decoder knows the filter coefficients in the GAE method (only), man could use the integer-to-integer

[4] Similar expressions can be obtained by replacing the indexes (m,n) with $(m,0)$ and $(m,n-1)$ with $(m-1,N_2-1)$; this corresponds to the change of row of the current pixel.

[5] We did not implement a fast version (see [30]) of the RLS algorithm.

Table 1
Algorithm of adaptation of the filter coefficients for the LAE method

Polyphase decomposition of the 2-D input signal into two signals x_1, x_2
Initialization
$\underline{c} = [1, 0, \ldots, 0]^T \in \mathbb{R}^{r_1+r_2}$
$\mathbf{\Phi}^{-1} = \delta^{-1}\mathbf{I}, 0 < \delta \ll 1.$
α, forgetting factor
For $m = 1, 2, \ldots$
For $n = 1, 2, \ldots$
make the vector $\underline{y}_1(m,n)$.
make the vector $\underline{y}_2(m,n)$.
make the observation vector $\underline{y}(m,n) = [\underline{y}_1^T(m,n), \underline{y}_2^T(m,n)]^T$
$x_\ell(m,n) = x_1(m,n)$
$x_h(m,n) = x_2(m,n) - \underline{c}^T \underline{y}(m,n)$
$\mathbf{\Phi}^{-1} = \alpha^{-1}\mathbf{\Phi}^{-1} - [\alpha^{-1}\mathbf{\Phi}^{-1}\underline{y}(m,n)\underline{y}(m,n)^T\mathbf{\Phi}^{-1}]/[\alpha + \underline{y}(m,n)^T\mathbf{\Phi}^{-1}\underline{y}(m,n)]$
$\underline{g}(m,n) = \mathbf{\Phi}^{-1}\underline{y}(m,n)$
$\underline{c} = \underline{c} + \underline{g}(m,n)x_h(m,n)$
end for n
end for m

generalized lifting scheme with GAE for a progressive coding in quality (or rate). Now it is not the case, because the criterion used does not take into account the distortion and is justified only for lossless coding. We emphasize that the problem for a progressive coding in rate is due to the choice of the adaptation criterion and not to the structure of the generalized lifting scheme.

4. Application to lossless coding of synthetic images

In this section we estimate the performances of one predict step (at the highest resolution) of the integer-to-integer generalized lifting scheme on artificial images that satisfy the assumptions of an optimal coding and we compare with other integer-to-integer decompositions.

4.1. Construction of the synthetic images

In our experiments, we used two families of Gaussian synthetic images, coded on 8 bits per pixel (bpp). The first one is composed of eight 2-D auto-regressive (AR) signals of size 512×512, which are each globally WSS. Half of these signals are generated with 2-D AR models which have a quarter-plan (QP) causality and the other half with a NSHP causality. They are shown in Fig. 6. The second family is a set of eight images of size 512×512, composed of several areas with different textures, which are each 2-D AR signals, simulating local stationarity (i.e., each area is stationary). The textures of the different areas have been generated with 2-D AR models having either QP or NSHP causality (chosen at random). They are shown in Fig. 6.

4.2. Experimental results

In this subsection, we compare for each synthetic image the variance and the first order entropy of subband x_h, obtained with the reversible version of the diagram in Fig. 3 (described in Section 2.4), with the GAE method ($p = 3, q = 3$, separable decimation), LAE method ($r_1 = 16$, $r_2 = 8$, $\alpha = 0.9995$, separable decimation), the reversible discrete wavelet decompositions (DWT) (9,7) and (5,3) of Daubechies and the method by Gerek and Çetin [11,31] mentioned above. Tables 2 and 3 present the performances of the different decompositions. For the two families of signals, the performances of the decompositions GAE and LAE are quite higher than those of the other decompositions and very close to one another. Indeed, on average, the gain is of 0.5 bpp for the globally WSS signals or 0.7 bpp for the locally stationary signals compared to the best of the other decompositions. We notice, as it was expected, that for globally stationary signals the GAE method gives slightly better results than LAE and that for locally stationary signals it is the LAE method which slightly precedes GAE. We also notice that the addition of the filter $B(z_1, z_2)$ appreciably improves (0.6 or 0.7 bpp on average) the performances of the adapted lifting scheme, thanks to a comparison with the method of Gerek and Çetin.

5. Comparative evaluations in lossless image compression

In this section, we compare the performances in lossless coding of the GAE and LAE methods with other well-known codecs: S + P with arithmetic coding [5], LOCO I [19] (i.e., the JPEG-LS standard), CALIC [21] and JASPER [23] (i.e., the JPEG2000 standard) on actual images.

5.1. Description of the images used in our tests

In our experiments, we considered four families of images: (1) 14 natural images of resolutions 512×512 and 1024×1024, (2) 14 satellite[6] images of various

[6] The satellite images have been given as a favor by the French National Center of Spatial Studies (CNES).

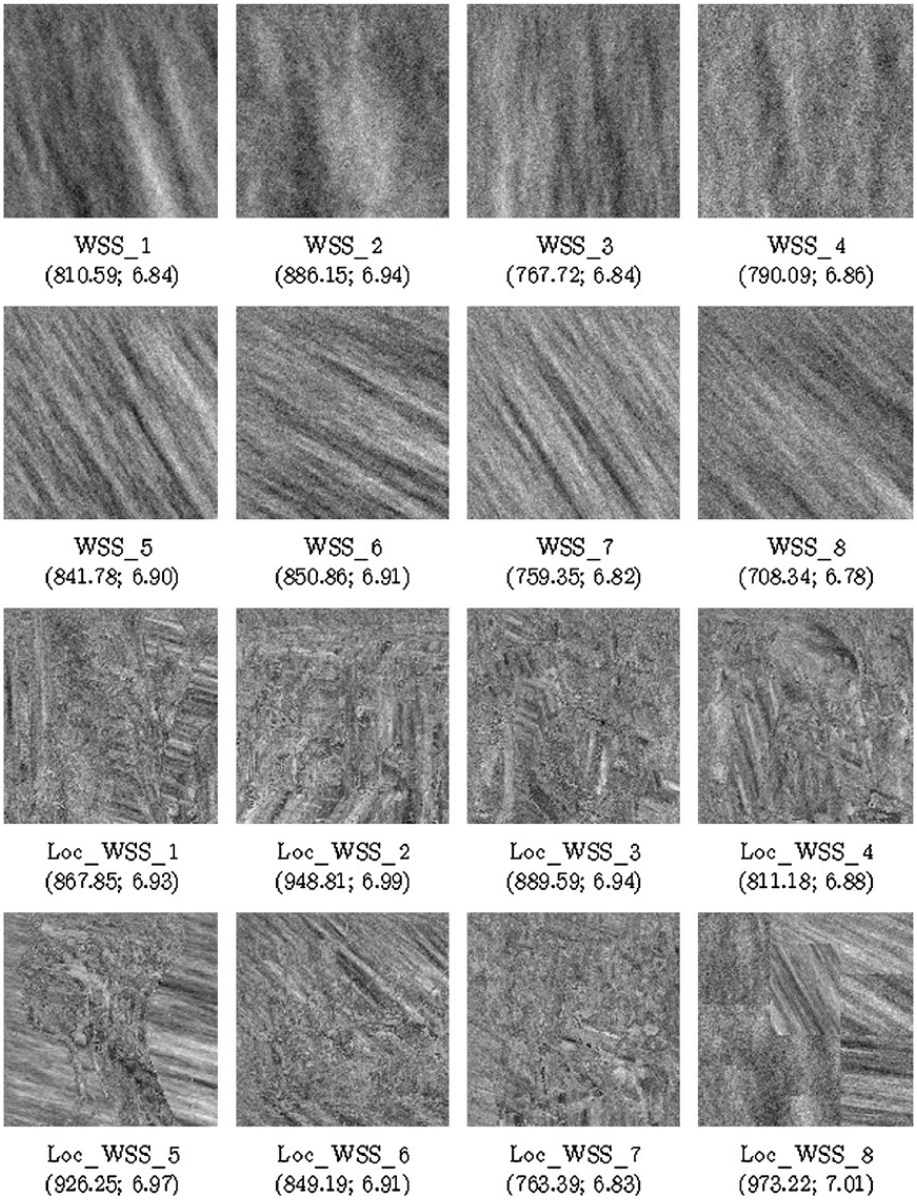


Fig. 6. Synthetic images, globally stationary (top first two rows) and locally stationary (last two rows). In the first row the causality is NSHP and QP in the second row. In the last two rows, the causality is either QP or NSHP per area (this choice has been made at random). The values indicated under each image correspond, respectively, to its variance and its first order entropy.

resolutions, (3) 16 MRI medical images of resolution 512×512 and (4) 11 images of textures and fingerprints whose resolution is either 512×512 or 768×768. All the images are coded with 8 bpp. In our tests, all the multiresolution decompositions applied to images of resolution 512×512 or 768×768 (resp. 1024×1024) use five (resp. six) levels of decomposition. Some images are shown in Fig. 7.

5.2. Description of the codecs associated with the LAE and GAE methods

For GAE and LAE methods, at each level of decomposition, meta parameters must be fixed, such as the type of twofold decimation (quincunx or separable), the orders of filters $A(z_1, z_2)$ and $B(z_1, z_2)$, the use of the S-transform or not (see Section 2.6) and, for LAE only, the value of the forgetting factor. For each family of images and for each level of decomposition, we looked for the best values of these meta parameters, i.e., the values that give in average the lowest first order entropy. We noted that these meta parameters vary from one family of images to another, whereas they slightly vary in the same family. Moreover, for the GAE method we noted that when the resolution is not greater than 64×64, in the adapted predict step, it is better to apply a process close to the LAR method [32] than to use the linear estimator described above, since there are not enough samples for a good estimation of the second order statistics. The process then consists in

Table 2
Variance (var.) and first order entropy (ent.) of the subband signal x_h for different lifting scheme decompositions

	GAE		LAE		Gerek et al.		(9,7)		(5,3)	
	Var.	Ent.	Var.	Ent.	Var.	Ent.	Var.	Ent.	Var.	Ent.
WSS_1	108.1	5.42	112.8	5.44	185.1	5.81	216.5	5.93	325.7	6.22
WSS_2	181.1	5.80	185.7	5.81	231.7	6.21	359.2	6.29	548.1	6.59
WSS_3	189.5	5.83	195.4	5.85	324.7	6.20	355.7	6.28	542.1	6.58
WSS_4	278.4	6.11	287.4	6.13	509.7	6.54	555.5	6.60	849.3	6.91
WSS_5	143.4	5.63	146.8	5.64	271.4	6.08	269.6	6.08	404.8	6.38
WSS_6	176.5	5.78	189.1	5.80	315.6	6.19	364.6	6.30	556.4	6.61
WSS_7	172.1	5.76	176.3	5.78	304.7	6.17	366.8	6.31	559.3	6.61
WSS_8	194.8	5.85	199.6	5.87	353.7	6.27	401.7	6.37	614.0	6.68
Average		5.77		5.79		6.18		6.27		6.57

The input signals WSS_1 to WSS_8 are described in Section 4.1.

Table 3
Variance (var.) and first order entropy (ent.) of the subband signal x_h for different lifting scheme decompositions

	GAE		LAE		Gerek et al.		(9,7)		(5,3)	
	Var.	Ent.	Var.	Ent.	Var.	Ent.	Var.	Ent.	Var.	Ent.
Loc_WSS_1	129.0	5.55	127.4	5.54	337.8	6.23	418.8	6.40	640.0	6.70
Loc_WSS_2	124.0	5.51	117.8	5.46	344.5	6.22	409.9	6.36	631.3	6.68
Loc_WSS_3	141.4	5.61	141.6	5.60	356.8	6.27	471.3	6.47	719.1	6.77
Loc_WSS_4	108.6	5.42	103.0	5.38	291.5	6.13	351.1	6.26	537.7	6.56
Loc_WSS_5	79.4	5.14	77.6	5.12	236.3	5.89	248.1	5.97	380.7	6.28
Loc_WSS_6	115.5	5.46	116.5	5.44	310.7	6.17	369.7	6.29	565.6	6.60
Loc_WSS_7	112.5	5.44	110.1	5.43	303.6	6.15	357.1	6.27	546.0	6.58
Loc_WSS_8	105.9	5.41	104.1	5.39	263.1	6.06	379.5	6.32	581.5	6.63
Average		5.44		5.42		6.14		6.29		6.60

The input signals Loc_WSS_1 to Loc_WSS_8 are described in Section 4.1.

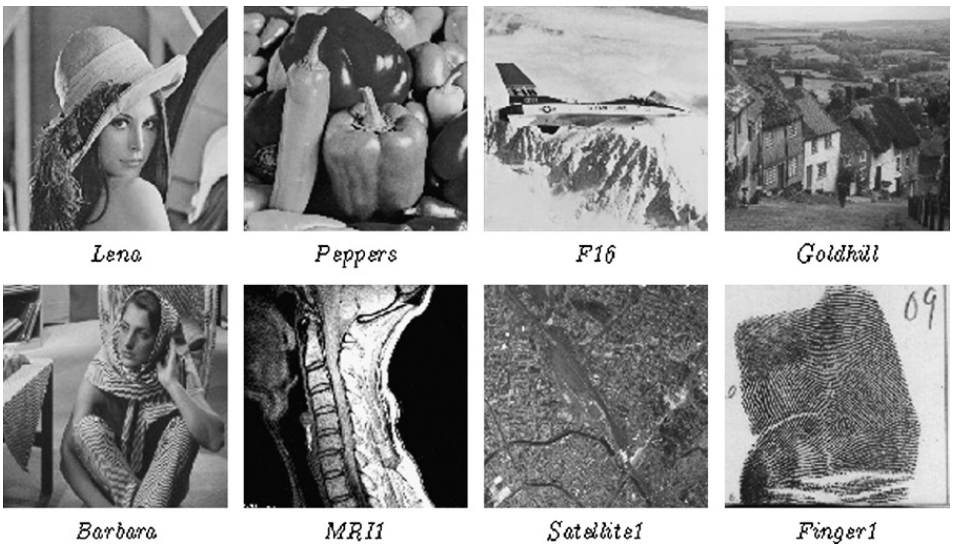


Fig. 7. Some images used in our tests.

dividing the image of the subband $x_1(m,n)$ in homogenous areas and in evaluating the estimation $\hat{x}_2(m,n)$ as the average of the pixels $x_1(m,n)$ which belong to the same homogenous area. The variant of GAE obtained by adapting the meta parameters to each family of images is noted GAEa in Table 5.

After the multi-resolution decomposition based on either the GAE or the LAE method, we applied a contextual adaptive arithmetic coder with one context per subband (we used the C sources S + P *image compression* by Said [5]). To rebuild the image, the decoder needs a heading in the bit stream containing the image size, the image mean,

the number of levels of decomposition and, for each level of decomposition, the orders of the filters, the filter coefficients (for GAE only), a Boolean specifying the type of down-sampling and another Boolean specifying whether the *S*-transform has been used or not. To reduce the size of the bit stream required for coding the heading, we truncated the filter coefficients (before applying the decomposition): a coefficient a is replaced by $\hat{a} = (\lfloor a.10^s + 0.5 \rfloor).10^{-s}$ (in our experiments, we chose $s = 6$) and the integer $10^s.\hat{a}$ is stored in the heading (for GAE only).

5.3. Experimental results and analysis

5.3.1. Mean first order entropy

We first compare the performances of the LAE and GAE decompositions with other integer-to-integer wavelet decompositions, without taking into account neither the bit stream required for coding the heading nor the entropic coder applied on the transformed coefficients. For this, we estimate the average of the first order entropy of the subbands, weighted by the ratio of the subband size on the entire image size. In order to be brief, the results of a few images are shown in Table 4, however, the row "Average" corresponds, for each family of images described in Section 5.1, to the average bit-rate of the whole family. The columns (a, b), where a and b are integers, correspond to integer-to-integer wavelets decompositions defined in [3]. In column S + P, the best predictor among A, B, and C of the transform S + P [5] has been chosen for each image.

We notice that for relatively smooth images (containing large areas of texture) like *Lena*, *Peppers*, *F16*, the average of first order entropies associated with GAE and LAE is only slightly lower than the one obtained with the S + P transformation. For images containing more outlines, like *Goldhill*, *Barbara* and the fingerprints *Finger*1–3, the proposed methods give the smallest average of first order entropies, the gain being approximatively 0.15 bpp.

Table 4
Average of the first order entropies of the image decomposed with 5 or 6 (images with an asterisk) levels of decomposition

	(2,2)	(4,2)	(4,4)	(5,3)	S + P	GAE	LAE
Lena	4.35	**4.30**	**4.30**	4.34	4.39	4.33	4.33
Goldhill	4.84	4.83	4.83	4.84	4.95	4.81	**4.80**
Barbara	4.99	4.86	4.82	4.99	4.83	4.69	**4.68**
F16	4.18	**4.14**	**4.14**	4.18	4.20	4.18	4.22
Mandrill	6.11	6.09	6.08	6.11	6.15	**6.07**	**6.07**
Peppers	**4.58**	**4.58**	**4.58**	**4.58**	4.70	**4.58**	4.67
IRM1	**2.31**	3.05	3.04	2.18	3.39	3.03	3.10
IRM2	2.35	3.10	3.10	**2.23**	3.49	2.78	3.05
IRM3	5.18	5.11	5.11	5.18	5.06	**4.88**	5.01
IRM4	4.54	4.47	4.48	4.52	**4.36**	4.43	4.38
IRM5	4.24	4.11	4.12	4.24	4.04	4.35	**4.03**
IRM6	1.84	2.36	2.36	**1.74**	2.60	3.47	2.39
Pentagone	5.27	5.27	**5.26**	5.27	5.72	5.27	5.28
Sanfrancisco	4.88	4.86	4.85	4.88	5.35	**4.85**	**4.85**
Okland	4.38	4.37	4.36	4.38	4.84	**4.35**	4.38
Toulouse	5.23	5.15	5.15	5.23	5.92	**5.05**	5.11
Genes	4.28	4.25	4.25	4.28	5.01	**4.21**	4.24
Airplane*	4.51	4.52	4.51	4.51	4.80	**4.46**	4.50
Airport*	5.40	5.38	5.38	5.40	5.82	**5.21**	5.31
Finger1	4.85	4.61	4.62	4.85	4.53	**4.33**	4.40
Finger2	4.25	4.09	4.10	4.25	4.09	**3.89**	3.94
Finger3	4.78	4.56	4.60	4.78	4.54	**4.36**	4.40

For the GAE method, the values of the parameters are $(p, q) = (3, 3)$ and a separable decimation. For the LAE method, the values of the parameters are separable decimation, $\alpha = 0.9995$, $(r_1, r_2) = (6, 2)$ for MRI medical images and $(r_1, r_2) = (8, 4)$ for all other image families. Bold indicates the lowest first order entropy of the row.

5.3.2. Actual bit-rate

We also compared the performances in lossless coding of the proposed methods on the actual bit-rate obtained with the contextual adaptive arithmetic coder implemented by Said [18]. Moreover, the bit-stream required for coding the heading is taken into account in the bit-rate. The results are shown in Table 5. For natural images, CALIC gives on average the smallest bit-rate, however, among codecs that permit progressive coding in resolution, the LAE method is that which gives the smallest bit-rate, slightly lower than the one of S + P or LOCO. Jasper is higher than other codecs by about 0.1 bpp. Whereas for MRI medical images, Jasper is significantly better. This is due to the fact that MRI medical images are smooth. The efficiency of Jasper decreases on images with steep outlines. On the family of textures and fingerprints, CALIC and LAE methods are similar and have significantly better performances than the other codecs.

On natural images and on MRI medical images, the performances in lossless coding of codecs based on the proposed methods do not exceed those of classical codecs based on multi-resolution decomposition with filters having fixed coefficients. However, as expected, the variance of the error of prediction $x_h(m, n)$ is smaller with the proposed methods than with the others. That illustrates the fact that the criterion of minimizing the variance for finding optimal transform in coding is only justified with Gaussian sources. For the families of satellite images and textures (with fingerprints), the codecs based on the proposed methods give a slight but still noticeable (about 0.05–0.08 bpp) coding gain compared to the others. However, the coding gain on satellite or textured images is much smaller than the one observed on synthetic images and we can deduce again that the criterion of minimizing the variance is not the good one for actual images.

5.3.3. Scalability and complexity

For a progressive coding in resolution, the *S*-transform is systematically applied in any "predict" step for resolutions smaller or equal to 256×256 (in order to avoid aliasing artifacts) and we observed that, compared to the others, the positive coding gain of codecs GAEa and LAE remains noticeable at smaller resolutions, for satellite and textured images. Now, as it was mentioned in Section 3.3, both the GAE and LAE methods are not suitable for progressive coding in rate (or quality). In Table 6 we compare the coding and decoding times of the different codecs.

Table 5
Bit-rates (in bpp) for the codecs S + P, LOCO (i.e., JPEG-LS), CALIC, Jasper (i.e., JPEG2000) and the proposed methods

Image	No progressive coding		Progressive coding in resolution			
	LOCO	CALIC	JASP	S + P	GAEa	LAE
Natural images						
Lena	4.24	4.13	4.32	**4.17**	4.26	4.21
Goldhill	4.71	4.65	4.84	**4.75**	4.78	**4.75**
Barbara	4.74	4.51	4.66	**4.53**	4.65	4.59
Mandrill	6.04	5.90	6.11	**5.93**	5.99	5.98
Peppers	4.49	4.39	4.62	**4.54**	4.58	4.56
Airplane	4.61	4.47	4.62	4.50	**4.45**	4.46
Airport	5.32	5.22	5.48	5.32	5.27	**5.24**
Average	4.82	4.70	4.91	4.78	4.78	**4.77**
MRI medical images						
MRI1	2.27	2.20	2.58	2.41	2.60	**2.40**
MRI2	2.54	2.34	**1.69**	2.59	2.31	2.13
MRI3	5.27	5.09	5.22	5.04	**5.02**	5.17
Average	2.81	2.65	**2.40**	2.86	2.71	2.54
Satellite images						
Genes	3.81	3.72	4.01	3.89	3.88	**3.85**
Mars	4.26	4.03	4.24	3.85	**3.76**	4.01
Okland3	4.41	4.28	4.44	4.31	**4.26**	4.28
Average	4.70	4.53	4.85	4.69	4.67	**4.61**
Textures and fingerprints images						
Fing1	4.57	4.44	4.46	4.33	4.25	**4.24**
Text1	6.71	6.62	6.79	6.53	6.48	**6.46**
Text2	5.97	5.88	6.18	5.95	**5.91**	5.93
Average	4.69	4.58	4.75	4.62	4.59	**4.57**

Only the four coders GAEa, LAE, Jasper and S + P allow for a progressive coding in resolution. The rows "average" are computed for each family of images. Bold indicates the lowest bit-rate of the row among the codecs that allow progressive resolution.

Table 6
Mean coding and mean decoding time on images of dimension 512 × 512 and coded on 8 bpp, expressed in seconds

Time	LOCO	CALIC	JASP	S + P	GAEa	LAE
Coding	0.12	0.21	0.40	0.36	3.19	3.27
Decoding	0.11	0.26	0.37	0.35	0.46	3.20

The codecs have been implemented on a PC PIII 700 MHz, with 256 Mo of RAM.

6. Conclusion

In this paper we have introduced an adapted generalized lifting scheme, in which the predict step is built upon two filters, leading to taking advantage of all the information available at the decoder. With this structure applied in a multi-resolution decomposition framework, we have studied two kinds of adaptation based on LSE, according to the different stationarity assumptions made on the input image. One decomposition, called globally adapted estimation (GAE), assumes the entire input image is a WSS signal. The other one, called locally adapted estimation (LAE), assumes only local WSS. The efficiency in lossless coding of these decompositions has been shown on Gaussian synthetic images satisfying these stationarity conditions and their performances have been compared with those of well-known codecs (S + P [18], LOCO I [20], CALIC [22] and Jasper [23]) on actual images. We have considered four families of images: natural, MRI medical, satellite and textures associated with fingerprints. On natural and MRI medical images, the performances in lossless coding of codecs based on the proposed methods do not exceed those of classical codecs. Nevertheless, for the families of satellite images and textures (with fingerprints), the codecs based on the proposed methods give a slight but still noticeable (about 0.05–0.08 bpp) coding gain compared to the others, at the price of a more important coding time. However, the coding gain for satellite and textured images is much smaller than the one observed on synthetic images and some improvements have to be done in order to satisfy the applications of satellite images. In future works we shall test other criteria, associated with the generalized lifting scheme, based on mutual information as the ones clarified in [33] or in [34].

Acknowledgments

The authors are grateful to Pierre Duhamel and anonymous reviewers for their very helpful comments as well as Jean-Luc Collette and Jean-Louis Gutzwiller who took part in the simulation work.

Appendix A. Fast calculation of the Yule–Walker equations

First we introduce new notations: for two vectors $\underline{u} \in \mathbb{R}^{\alpha}$ and $\underline{v} \in \mathbb{R}^{\beta}$, we denote $\mathbf{T}[\underline{u}, \underline{v}]$ the Toeplitz matrix of dimension $\alpha \times \beta$ whose first column (resp. row) is equal to $\underline{u}$ (resp. $\underline{v}^{\mathrm{T}}$). For a matrix $\mathbf{A}$ of dimension $m \times n$, the element localized at the intersection of the $(i+1)$th row and the $(j+1)$th column ($0 \leqslant i < m$ and $0 \leqslant j < n$) is denoted by $[\mathbf{A}]_{i,j}$. Let us introduce the vector $\underline{b}_0 = (0, \ldots, 0, -1, b_{0,1}, b_{0,2}, \ldots, b_{0,q})^{\mathrm{T}}$ of dimension $2q+1$, the p vectors $\underline{b}_{\ell} = (b_{\ell,-q}, b_{\ell,-q+1}, \ldots, b_{\ell,q})^{\mathrm{T}}$ ($1 \leqslant \ell \leqslant p$) of dimension $2q+1$, the $2p+1$ vectors $\underline{a}_{\ell} = (a_{\ell,-q}, a_{\ell,-q+1}, \ldots, a_{\ell,q})^{\mathrm{T}}$ ($-p \leqslant \ell \leqslant p$) of dimension $2q+1$ and the vectors $\underline{b}' = (\underline{b}_0^{\mathrm{T}}, \underline{b}_1^{\mathrm{T}}, \ldots, \underline{b}_p^{\mathrm{T}})^{\mathrm{T}}$ and $\underline{a} = (\underline{a}_{-p}^{\mathrm{T}}, \underline{a}_{-p+1}^{\mathrm{T}}, \ldots, \underline{a}_p^{\mathrm{T}})^{\mathrm{T}}$. With the $(m+1)$th row of the subband signal x_2, we associate the $p+1$ Toeplitz matrices ($0 \leqslant \ell \leqslant p$)

$$\begin{aligned}\mathbf{X}_{\ell}(m) = \mathbf{T}[&(x_2(m-\ell, q), \ldots, x_2(m-\ell, N_2-1))^{\mathrm{T}},\\ &(x_2(m-\ell, q), \ldots, x_2(m-\ell, 0),\\ &0, \ldots, 0)^{\mathrm{T}}]\end{aligned} \tag{23}$$

of dimension $(N_2 - q) \times (2q+1)$ and the block matrix

$$\mathbf{X}(m) = [\mathbf{X}_0(m)\ \mathbf{X}_1(m)\ \cdots\ \mathbf{X}_p(m)] \tag{24}$$

of dimension $(N_2 - q) \times (2q+1)(p+1)$. In the same way, with the $(m+1)$th row of the signal x_1 we associate the

$2p+1$ Toeplitz matrices ($-p \leqslant \ell \leqslant p$)

$$\mathbf{Y}_\ell(m) = \mathbf{T}[(x_1(m-\ell,q),\ldots,x_1(m-\ell,N_2-1))^{\mathrm{T}}, (x_1(m-\ell,q),\ldots,x_1(m-\ell,0), 0,\ldots,0)^{\mathrm{T}}] \quad (25)$$

of dimension $(N_2-q)\times(2q+1)$ and the block matrix $\mathbf{Y}(m)=[\mathbf{Y}_{-p}(m)\mathbf{Y}_{-p+1}(m)\cdots\mathbf{Y}_p(m)]$ of dimension $(N_2-q)\times(2p+1)(2q+1)$. With these notations, the error of estimation $\underline{x}_h(m)=[x_h(m,0),\ldots,x_h(m,N_2-q-1)]^{\mathrm{T}}$ associated to the $(m+1)$th row of x_2 satisfies the relation $-\underline{x}_h(m)=\mathbf{X}(m)\underline{b}'+\mathbf{Y}(m)\,\underline{a}$. Moreover, it results from relations (23) and (25) that $\mathbf{Y}_k(m)=\mathbf{Y}_{k-1}(m-1)=\mathbf{Y}_0(m-k)$, $\mathbf{X}_k(m)=\mathbf{Y}_k(m)=\mathbf{0}$ (if $m<k$) and that for $0<k\leqslant p$ and $0\leqslant m<M_2-p$

$$\mathbf{X}_k(m)=\mathbf{X}_{k-1}(m-1)=\mathbf{X}_0(m-k). \quad (26)$$

The equations of estimation lead then to

$$-(\underline{x}_h(0)^{\mathrm{T}},\underline{x}_h(1)^{\mathrm{T}},\ldots,\underline{x}_h(M_2-p-1)^{\mathrm{T}})^{\mathrm{T}}=\mathbf{X}\underline{b}'+\mathbf{Y}\,\underline{a} \quad (27)$$

with the block Toeplitz matrices $\mathbf{X}$ and $\mathbf{Y}$, whose first columns are, respectively, $(\mathbf{X}_0(0)^{\mathrm{T}},\mathbf{X}_0(1)^{\mathrm{T}},\ldots,\mathbf{X}_0(M_2-p-1)^{\mathrm{T}})^{\mathrm{T}}$ and $(\mathbf{Y}_0(p)^{\mathrm{T}},\mathbf{Y}_0(p+1)^{\mathrm{T}},\ldots,\mathbf{Y}_0(M_2-1)^{\mathrm{T}})^{\mathrm{T}}$ and whose first rows are, respectively, $(\mathbf{X}_0(0),\mathbf{0},\ldots,\mathbf{0})$ and $(\mathbf{Y}_0(p),\mathbf{Y}_0(p-1),\ldots,\mathbf{Y}_0(0),\mathbf{0},\ldots,\mathbf{0})$. Since each block $\mathbf{X}_0(k)$ or $\mathbf{Y}_0(k)$ is Toeplitz, we may observe that both the matrices $\mathbf{X}$ and $\mathbf{Y}$ are Toeplitz and also block Toeplitz. The coefficients of the optimal filters are solutions to the Yule–Walker equations [28] that may be written $(\underline{b}'^{\mathrm{T}},\underline{a}^{\mathrm{T}})\mathbf{C}=(\underline{u}^{\mathrm{T}},-W,\underline{0}^{\mathrm{T}})$, where $\underline{u}$ is a vector of dimension q, $W=(M_2-p)(N_2-q)J_1$ and

$$\mathbf{C}=[\mathbf{X}\ \mathbf{Y}]^{\mathrm{T}}[\mathbf{X}\ \mathbf{Y}]=\begin{bmatrix}\mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{Y}\\ \mathbf{Y}^{\mathrm{T}}\mathbf{X} & \mathbf{Y}^{\mathrm{T}}\mathbf{Y}\end{bmatrix} \quad (28)$$

is a symmetrical matrix of order $\omega'=(2q+1)(3p+2)$. Since the $q+1$ first components of $\underline{b}'$ are known, the actual dimension of the system is $\omega=6pq+3p+3q+1$. However, it is faster to calculate first the entire matrix $\mathbf{C}$ and then to extract from it the system's matrix $\mathscr{Y}^{\mathrm{T}}\mathscr{Y}$ than to directly calculate this last matrix. Relation (27) differs from the well-known equations encountered with the prewindowed method [27,28] and from the equations in [10] in the extra terms $\mathbf{Y}\,\underline{a}$. Consequently, the displacement rank [27] of the matrix $\mathbf{C}$ is not equal to 1. Nevertheless, the elements of $\mathbf{C}$ still have a great redundancy and the reasoning presented in [29] can be adapted to this matrix.

To be short, only the relations that permit to calculate the block $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ of $\mathbf{C}$ are given. The same reasoning could be carried out for the other blocks. It results from relations (27) and (24) that $\mathbf{X}^{\mathrm{T}}\mathbf{X}=\sum_{m=0}^{M_2-p-1}\mathbf{X}(m)^{\mathrm{T}}\mathbf{X}(m)=[\sum_{m=0}^{M_2-p-1}\mathbf{X}_k(m)^{\mathrm{T}}\mathbf{X}_\ell(m)]$ (for $0\leqslant k,\ell<p$), where the last expression is a block representation of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Let $\mathbf{T}_{k,\ell}=\sum_{m=0}^{M_2-p-1}\mathbf{X}_k(m)^{\mathrm{T}}\mathbf{X}_\ell(m)$ ($0\leqslant k,\ell<p$) be the block of dimension $(2q+1)\times(2q+1)$. We have $\mathbf{T}_{k,\ell}^{\mathrm{T}}=\mathbf{T}_{\ell,k}$. Relation (26) leads to $\mathbf{T}_{k,\ell}=\mathbf{T}_{k-1,\ell-1}-\mathbf{X}_0(M_2-p-k)^{\mathrm{T}}\mathbf{X}_0(M_2-p-\ell)$ for $1\leqslant k\leqslant\ell\leqslant p$. It is straightforward to deduce from (23) that $\mathbf{X}_0(m-k)^{\mathrm{T}}\mathbf{X}_0(m-\ell)$ ($0\leqslant k\leqslant\ell\leqslant p$ and $\ell\leqslant m<M_2-p$) has a displacement rank of 2. Therefore, for $0\leqslant\ell\leqslant p$, we have $[\mathbf{T}_{0,\ell}]_{i,j}=[\mathbf{T}_{0,\ell}]_{i-1,j-1}+\sum_{m=\ell}^{M_2-p-1}x_2(m,q-i)x_2(m-\ell,q-j)-\sum_{m=\ell}^{M_2-p-1}x_2(m,N_2-i)x_2(m-\ell,N_2-j)$. In conclusion, the calculation of a block requires (in additions and multiplications):

- $O(N_2)$ operations for $\mathbf{T}_{k,\ell}$ with $0\leqslant k\leqslant\ell\leqslant p$,
- $2(4q+1)M_2N_2+o(M_2N_2)$ operations for $\mathbf{T}_{0,\ell}$ with $p\geqslant\ell\geqslant1$,
- $2(2q+1)M_2N_2+o(M_2N_2)$ operations for $\mathbf{T}_{0,0}$,

and the calculus of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ requires $2[(4q+1)p+2q+1]M_2N_2+o(M_2N_2)$ operations. In the same way, the computation of $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ requires $2(4q+1)(3p+1)+o(M_2N_2)$ operations and the computation of $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ requires $2[(4q+1)2p+2q+1]M_2N_2+o(M_2N_2)$ operations. Finally, the computation of $\mathbf{C}$ costs $2[2(4q+1)(3p+1)+1]M_2N_2+o(M_2N_2)$ operations.

References

[1] M. Antonini, M. Barlaud, P. Mathieu, I. Daubechies, Image coding using wavelet transform, IEEE Trans. Image Process. 1 (2) (April 1992) 205–219.

[2] W. Sweldens, The lifting scheme: a new philosophy in biorthogonal wavelet construction, Proc. SPIE 2569 (September 1995) 68–78.

[3] A.R. Calderbank, I. Daubechies, W. Sweldens, B.L. Yeo, Wavelet transforms that map integers to integers, Appl. Comput. Harmon. Anal. 3 (2) (1996) 186–200 IEEE Trans. Image Process. 10 (1) (January 2001) 1–14.

[4] I. Daubechies, W. Sweldens, Factoring wavelet transforms into lifting steps, J. Fourier Anal. Appl. 4 (3) (1998) 247–269.

[5] A. Said, W.A. Pearlman, An image multiresolution representation for lossless and lossy compression, IEEE Trans. Image Process. 5 (September 1996) 1303–1310.

[6] M.D. Adams, F. Kossentini, Reversible integer-to-integer wavelet transforms for image compression: performance evaluation and analysis, IEEE Trans. Image Process. 9 (6) (June 2000) 1010–1024.

[7] F.J. Hampson, J.C. Pesquet, *M*-band nonlinear subband decomposition with perfect decomposition, IEEE Trans. Image Process. 7 (11) (November 1998) 1547–1560.

[8] I. Amonou, P. Duhamel, Nonredundant representation of images allowing object-based and multiresolution scalable coding, in: Proceedings of the International Conference on Visual Communications and Image Processing, Perth, Australia, June 2000, pp. 598–608.

[9] Y.-S. Chung, M. Kanefsky, On 2D recursive LMS algorithms using ARMA prediction for ADPCM encoding of images, IEEE Trans. Image Process. 1 (3) (July 1992) 416–422.

[10] X. Wu, K.U. Barthel, W. Zhang, Piecewise 2D autoregression for predictive image coding, in: Proceedings of the IEEE International Conference on Image Processing, vol. 3, Chicago, USA, October 1998, pp. 901–904.

[11] Ö.N. Gerek, A.E. Çetin, Adaptive polyphase subband decomposition structures for image compression, IEEE Trans. Image Process. 9 (10) (October 2000) 1649–1660.

[12] N.V. Boulgouris, D. Tzovaras, M.G. Strintzis, Lossless image compression based on optimal prediction, adaptive lifting, and conditional arithmetic coding, IEEE Trans. Image Process. 10 (1) (January 2001) 1–14.

[13] R.L. Claypoole, G.M. Davis, W. Sweldens, R. Baraniuk, Nonlinear wavelet transform for image coding via lifting, IEEE Trans. Image Process. 2 (12) (December 2003) 1449–1459.

[14] A. Gouze, A. Antonini, M. Barlaud, B. Macq, Optimized lifting scheme for two-dimensional quincunx sampling, in: Proceedings of the IEEE International Conference on Image Processing, vol. 2, Thessaloniki, Greece, October 2001, pp. 253–256.

[15] M. Barret, H. Bekkouche, Adapted nonlinear multiresolution decomposition with applications in progressive lossless image coding, in: Proceedings of the International Symposium on Image

and Signal Processing and Analysis, Pula, Croatia, June 2001, pp. 609–613.
[16] H. Bekkouche, M. Barret, Adaptive multiresolution decomposition: application to lossless image compression, in: Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing, Orlando, FL, USA, May 2002.
[17] H. Bekkouche, M. Barret, Comparison of lossless codecs for satellite and MRI medical images, in: Proceedings of the XI European Signal Processing Conference, Toulouse, France, September 2002.
[18] ⟨http://www.cipr.rpi.edu/research/SPIHT/EW_Code/lossless.zip⟩.
[19] M.J. Weinberger, G. Seroussi, G. Sapiro, The LOCO-I lossless image compression algorithm: principles and standardization into JPEG-LS, IEEE Trans. Image Process. 9 (August 2000) 1309–1324.
[20] ⟨http://www.hpl.hp.com/loco/software.htm⟩.
[21] X. Wu, N. Memon, Context-based, adaptive, lossless image coding, IEEE Trans. Comm. 45 (4) (April 1997) 437–444.
[22] ⟨http://www.csd.uwo.ca/faculty/wu/⟩.
[23] ⟨http://www.ece.ubc.ca/~mdadams/jasper/⟩.
[24] W. Sweldens, The lifting scheme: a custom-design construction of biorthogonal wavelets, Appl. Comput. Harmon. Anal. 3 (2) (1996) 186–200.
[25] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, New York, 1999.
[26] M. Barret, in: M. Najim (Ed.), Digital Filters Design for Signal and Image Processing, ISTE Ltd., 2006 (Chapter 10: Filter stability and Chapter 11: The two-dimensional domain).
[27] B. Friedlander, M. Morf, T. Kailath, L. Ljung, New inversion formulas for matrices classified in terms of their distance from Toeplitz matrices, Linear Algebra Appl. 27 (1979) 31–60.
[28] S. Haykin, Adaptive Filter Theory, second ed., Prentice-Hall, Englewood Cliffs, NJ, 1991.
[29] T. Kailath, S.Y. Kung, M. Morf, Displacement ranks of a matrix, Bull. Amer. Math. Soc. 1 (5) (September 1979) 769–773.
[30] X. Liu, M. Najim, A two-dimensional fast lattice recursive least squares algorithm, IEEE Trans. Signal Process. 44 (10) (October 1996).
[31] Ö.N. Gerek, A.E. Çetin, Polyphase adaptive filter banks for fingerprint image compression, in: Proceedings of the 9th European Signal Processing Conference, Rhodes, Grèce, September 1998, pp. 45–48.
[32] O. Déforges, J. Ronsin, Locally adaptive resolution method for progressive still image coding, in: Proceedings of the International Symposium on Signal Processing and its Applications, Brisbane, Australia, 22–25 August 1999.
[33] M. Narozny, M. Barret, D.T. Pham, ICA based algorithms for computing optimal 1-D linear block transforms in variable high-rate source coding, Signal Processing 88 (2) (2008) 268–283.
[34] A. Benazza-Benyahia, J.C. Pesquet, J. Hattay, H. Masmoudi, Block-based adaptive vector lifting schemes for multichannel image coding, EURASIP J. Image Video Process. 2007 (2007) 10 article ID 13421.

Available online at www.sciencedirect.com
ScienceDirect
Signal Processing 88 (2008) 268–283
SIGNAL PROCESSING
www.elsevier.com/locate/sigpro

ICA based algorithms for computing optimal 1-D linear block transforms in variable high-rate source coding ☆

Michel Narozny[a,*], Michel Barret[a], Dinh-Tuan Pham[b]

[a] *SUPELEC, Information Multimodality & Signal Team, 2 rue É. Belin 57070 Metz, France*
[b] *Jean Kuntzmann Laboratory, 51 rue des Mathématiques, BP 53, 38041 Grenoble Cedex 9, France*

Received 27 June 2006; received in revised form 19 June 2007; accepted 17 July 2007
Available online 26 July 2007

Abstract

The Karhunen–Loève Transform (KLT) is optimal for transform coding of Gaussian sources, however, it is not optimal, in general, for non-Gaussian sources. Furthermore, under the high-resolution quantization hypothesis, nearly everything is known about the performance of a transform coding system with entropy constrained scalar quantization and mean-square distortion. It is then straightforward to find a criterion that, when minimized, gives the optimal linear transform under the abovementioned conditions. However, the optimal transform computation is generally considered as a difficult task and the Gaussian assumption is then used in order to simplify the calculus. In this paper, we present the abovementioned criterion as a contrast of independent component analysis modified by an additional term which is a penalty to non-orthogonality. Then we adapt the `icainf` algorithm by Pham in order to compute the transform minimizing the criterion either with no constraint or with the orthogonality constraint. Finally, experimental results show that the transforms we introduced can (1) outperform the KLT on synthetic signals, (2) achieve slightly better PSNR for high-rates and better visual quality (preservation of lines and contours) for medium-to-low rates than the KLT and 2-D DCT on grayscale natural images.

Keywords: Transform coding; Independent component analysis; Non-Gaussian signals; Image compression

1. Introduction

Transform coding has been extensively developed for coding natural signals like sounds, images (e.g., JPEG [1]) and video. The motivating principle of transform coding is that *simple* coding may be more effective in the transform domain than in the original signal space. Generally "simple coding" corresponds to the use of scalar quantization and scalar entropy coding, because they provide a good trade-off between computational complexity and performance [2].

In a conventional transform coder [3,4], an input vector $\mathbf{X}$ is first transformed into another vector $\mathbf{Y} = \mathbf{TX}$ of the same dimension. The components of that vector are then described to the decoder using independent scalar quantizers on the coefficients. Finally, the decoder reconstructs the quantized

☆ This work was supported in part by the French Ministry of Higher Education and Research, with the "Action Concertée Incitative Masse de Données" ACI^{2}M project, and * by the Lorraine Region.
*Corresponding author. Tel.: +33 612 27 07 57.
E-mail addresses: Michel.Narozny@supelec.fr (M. Narozny), Michel.Barret@supelec.fr (M. Barret), Dinh-Tuan.Pham@imag.fr (D.-T. Pham).

0165-1684/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2007.07.017

transform vector $\widehat{\mathbf{Y}}$ and then uses a linear transformation $\mathbf{U}$ to get an estimate of the original input vector $\widehat{\mathbf{X}} = \mathbf{U}\widehat{\mathbf{Y}}$. In this paper, we consider mean-square distortion : $\mathrm{E}[\|\mathbf{X} - \widehat{\mathbf{X}}\|^2]$, where E denotes mathematical expectation. When the input vector is Gaussian, the optimal transforms $\mathbf{T}$ and $\mathbf{U}$ satisfy both the conditions $\mathbf{T}$ is an orthogonal matrix (i.e., $\mathbf{T}^{-1} = \mathbf{T}^{\mathrm{T}}$, where $^{\mathrm{T}}$ denotes transposition) that produces uncorrelated transform coefficients and $\mathbf{U} = \mathbf{T}^{-1}$. Such a transform $\mathbf{T}$ is often called Karhunen–Loève transform (KLT) in coding lessons.

The optimality of the KLT and its inverse is well known for high rates [3], or when optimal fixed-rate quantizers are employed [5], or more generally when the quantizers are scale-invariant [2] for both the fixed-rate and the variable-rate coding models. Moreover, for non-Gaussian sources, it is well known that the KLT can be suboptimal in transform coding [6]. Research in transform coding theory, including subband coding, has been constant for several decades. There is a great interest in extending theoretical results on transform optimality towards low bit-rates [7,2], or in extending the KLT optimality in more general situations by weakening the assumptions that ensure the optimality. In [8], it is shown that under the classical assumptions of high-rate quantization, mean-square distortion and variable-rate coding, the KLT remains optimal for a more general class of signals than Gaussian ones. In [9], the quantization effect on backward adaptive transform coding of Gaussian sources is studied; the perturbation due to quantization on coding gain is computed up to second order for KLT and causal unitary transform.

Under the high-resolution quantization hypothesis, nearly everything is known about the performance of a transform coding system with entropy constrained scalar quantization and mean-square distortion. It is then straightforward to find a criterion that, when minimized, gives the optimal linear transform under the abovementioned conditions. Nevertheless, the optimal transform computation is generally considered as a difficult task [7] and the Gaussian assumption is then used in order to simplify the calculus. In this paper we resolve the problem of computing the optimal transform under high-rate entropy constraint scalar quantization hypothesis.

In 1963, Huang and Schultheiss [5] showed that, for scalar Lloyd–Max quantizers, if the vector source $\mathbf{X}$ is Gaussian then the mean-square distortion is minimized by choosing $\mathbf{U} = \mathbf{T}^{-1}$ when $\mathbf{T}$ is a KLT of $\mathbf{X}$. Actually, in their proof, they used on the one hand the fact that, applied to a real random variable Y, a Lloyd–Max scalar quantizer with cells $S_1, \ldots, S_k$ and outputs $\widehat{y}_1, \ldots, \widehat{y}_k$ satisfies the *centroid condition* $\widehat{y}_i = \mathrm{E}[Y | Y \in S_i]$ for $(1 \leqslant i \leqslant k)$; and in the other hand, the following property satisfied by Gaussian sources when $\mathbf{T}$ is a KLT: the jth component of the quantization noise $\mathbf{Y} - \widehat{\mathbf{Y}}$ and the ith component of $\widehat{\mathbf{Y}}$ are uncorrelated for $i \neq j$. When the Gaussian assumption does not hold, it is clear that the previous property is satisfied if the components of $\mathbf{Y}$ are statistically independent (indeed $Y_j - \widehat{Y}_j$ and $\widehat{Y}_i$ are deterministic functions of Y_j and Y_i, respectively). Hence their proof remains valid under this condition. Moreover, the result by Huang and Schulteiss still holds when the Lloyd–Max scalar quantizers are replaced by uniform scalar quantizers, under the additional assumption of high-rate quantization [3] (since they satisfy the centroid condition). In this paper, we suppose the transform used by the decoder satisfies $\mathbf{U} = \mathbf{T}^{-1}$. This assumption is partially justified by the fact that the linear transforms $\mathbf{T}$ we consider in the following give transformed vectors $\mathbf{Y}$ with minimal mutual information between components, and hence with components generally close to independence. However, the total independence is rarely achieved and it will be interesting to complete our study by the general case of arbitrary invertible matrices $\mathbf{T}$ and $\mathbf{U}$.

In this paper, we first show that under the high-rate entropy constraint scalar quantization hypothesis, for mean-square distortion and the condition $\mathbf{U} = \mathbf{T}^{-1}$, the criterion that gives—when minimized—the optimal linear transform $\mathbf{T}$ can be expressed as a classical contrast of independent component analysis (ICA) (actually the opposite of such a contrast), modified by an additional term which can be explained as a pseudo-distance to orthogonality. This new presentation of a classical result gives an interesting point of view covering ICA and data compression. Indeed, it is well known that images of natural scenes are well modeled with ICA [10], this underlies the potential usefulness of ICA to compression [11–13]. Then, we present two variants of the `icainf` algorithm by Pham [14] which can compute the optimal transform $\mathbf{T}$ with (1) no constraint and (2) the orthogonality constraint. Finally, experimental results on both synthetic data and natural images are given in Section 4. They show that the transforms returned by the new

algorithms can (1) outperform the KLT when used with synthetic signals and (2) achieve slightly better performance for high-rates and better visual quality (preservation of lines and contours) than the KLT and 2-D DCT when applied to the medium-to-low bit-rate compression of natural images. This last result is unexpected, since no optimality is ensured at low bit-rates. The paper begins with a brief review of high bit-rate transform coding. The results presented here have been partially published in [15,16].

2. High bit-rate transform coding

The general structure of a transform coding scheme is shown in Fig. 1. The class of signals to be encoded is represented by a random vector $\mathbf{X} = [X_1, \ldots, X_N]^{\mathrm{T}}$ of size N. Although these signals may be multidimensional like images, they are indexed by an integer to simplify notations: $\mathbf{X}(m)$. The components of $\mathbf{X}(m)$ are successive samples from a source signal $(x(n))_n$. A conventional transform coder applies a linear invertible transform $\mathbf{T}: \mathbb{R}^N \to \mathbb{R}^N$ to $\mathbf{X}$ in order to obtain a random vector $\mathbf{Y} = [Y_1, \ldots, Y_N]^{\mathrm{T}}$ better suited to coding than $\mathbf{X}$. To construct a finite code, each coefficient Y_i is first approximated by a quantized value $\widehat{Y}_i$. We concentrate only on scalar quantizers. The quantized coefficients are then entropy coded. The coded representation is stored or communicated over an error-corrected (lossless) channel. The receiver (decoder) provides an approximation $\widehat{\mathbf{X}} = [\widehat{X}_1, \ldots, \widehat{X}_N]^{\mathrm{T}}$ of the original signal $\mathbf{X}$ by applying a linear transform $\mathbf{U}$ to the quantized signal $\widehat{\mathbf{Y}}$. In this paper we assume $\mathbf{U} = \mathbf{T}^{-1}$.

A relevant problem for a system designer is to minimize the reconstruction error under bit-rate constraint. Here, the optimization criterion is the mean-square distortion $D = \frac{1}{N}\mathrm{E}[\|\mathbf{X} - \widehat{\mathbf{X}}\|^2]$ which satisfies

$$D = \frac{1}{N}\sum_{i=1}^{N} w_i E[(Y_i - \widehat{Y}_i)^2] + \frac{1}{N}\sum_{i=2}^{N}\sum_{j=1}^{i-1} w_{ij} E[(Y_i - \widehat{Y}_i)(Y_j - \widehat{Y}_j)], \quad (1)$$

where $w_i = \sum_{j=1}^{N} [\mathbf{T}^{-1}]_{j,i}^2$ $(1 \leqslant i \leqslant N)$, $w_{ij} = \sum_{k=1}^{N} 2[\mathbf{T}^{-1}]_{k,i}[\mathbf{T}^{-1}]_{k,j}$ $(1 \leqslant j \leqslant i \leqslant N)$, and $[\mathbf{A}]_{i,j}$ denotes the element localized on the ith row and the jth column of $\mathbf{A}$. The last term in (1) vanishes when (1) $w_{ij} = 0$ for $i \neq j$, i.e., the transform $\mathbf{T}$ is orthogonal—or more generally when the column vectors of $\mathbf{T}^{-1}$ are pairwise orthogonal—or (2) when the quantization noises of different quantizers are uncorrelated and centered. The last condition (2) is satisfied when (i) the vector $\mathbf{X}$ is Gaussian and the transform $\mathbf{T}$ is a KLT[1] or when (ii) the transform coefficients are statistically independent—the Gaussian assumption is then useless—and the quantizers satisfy the centroid condition. Moreover, experimental tests show that condition (2) approximatively holds under high-rate quantization hypothesis for any kind of signal.

In the following, we assume that the end-to-end distortion D can be well approximated by a weighting sum of the distortion $D_i = \mathrm{E}[(Y_i - \widehat{Y}_i)^2]$ of each transform coefficient as

$$D \approx \frac{1}{N}\sum_{i=1}^{N} w_i D_i. \quad (2)$$

Approximation (2) is an equality when $\mathbf{T}$ is orthogonal or when $\mathbf{Y}$ has independent components; it is a good approximation for any $\mathbf{T}$ when the transform coefficients are close to independence and last it is a good approximation for any $\mathbf{T}$ under the high-rate quantization hypothesis.

2.1. Entropy-constrained scalar quantization

A scalar quantizer Q approximates a random variable Y by a quantized variable $\widehat{Y} = Q(Y)$. Q is a mapping from a source alphabet $\mathbb{R}$ to a reproduction codebook $\mathscr{C} = \{\widehat{y}_i\}_{i \in \mathscr{K}} \subset \mathbb{R}$, where $\mathscr{K}$ is an arbitrary countable index set. We denote $p_i = P\{\widehat{Y} = \widehat{y}_i\}$. The Shannon theorem [3] proves that the entropy $H(\widehat{Y}) = -\sum_i p_i \log_2 p_i$ is a lower bound on the average number of bits per symbol used to encode the values of $\widehat{Y}$. Arithmetic entropy coding [17,18] achieves an average bit-rate that can be arbitrarily close to the entropy lower bound; therefore, we shall consider that this lower bound is reached. An *entropy constrained scalar quantizer* is designed to minimize $H(\widehat{Y})$ for a fixed mean-square distortion $D = \mathrm{E}[(Y - \widehat{Y})^2]$. It is well known [3] that for a fixed distortion D, under the high-resolution quantization hypothesis and if we assume that the random variable Y admits

[1] Indeed, the transform coefficients $(Y_1, \ldots, Y_N)$ are then independent and hence the quantization noises $(Y_1 - \widehat{Y}_1, \ldots, Y_N - \widehat{Y}_N)$ which are deterministic functions of the transform coefficients are uncorrelated, moreover they are centered when the quantizers satisfy the centroid condition.

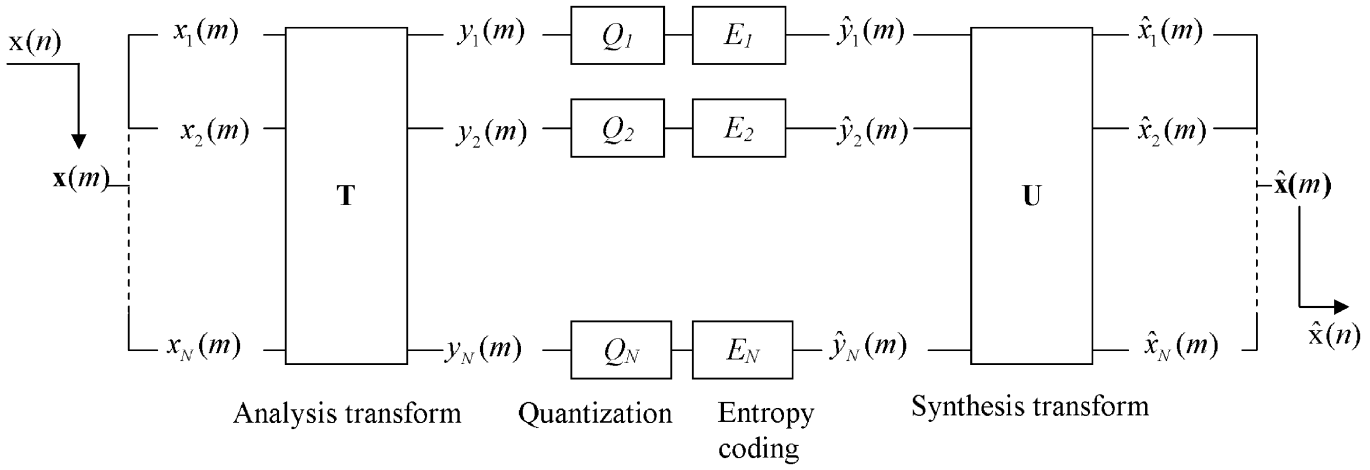


Fig. 1. Compression system including a linear transform, a quantization and an entropy coding stage.

a probability density function (pdf) $p(y)$, then the minimum average bit-rate $R = H(\widehat{Y})$ is achieved by a uniform quantizer, and $R \approx h(Y) - \frac{1}{2}\log_2(12D)$, where $h(Y) = \int \log_2[p(y)]p(y)\,dy$ is the differential entropy of Y. Generally it is preferable to introduce the variance σ^2 of Y and the differential entropy $h(\tilde{Y})$ of the standardized random variable $\tilde{Y} = (Y - \mathrm{E}[Y])/\sigma$: $h(\tilde{Y}) = h(Y) - \log_2 \sigma$ in order to separate the contribution of the signal power with that of its pdf shape (of course, this is possible only if Y admits finite second-order statistics). The distortion rate satisfies then

$$D \approx c\sigma^2 2^{-2R} \quad \text{with } c = \frac{2^{2h(\tilde{Y})}}{12}, \tag{3}$$

where the constant c depends only on the pdf shape.

2.2. Optimal bit allocation

Coding (quantizing and entropy coding) each transform coefficient Y_i separately splits the total number of bits among the transform coefficients in some manner. This bit allocation problem can be stated this way: one is given a set of quantizers described by their distortion-rate performances as $D_i \approx c_i\sigma_i^2 2^{-2R_i}$, $R_i \in \mathscr{R}_i$ for $(1 \leqslant i \leqslant N)$. Each set of available rates $\mathscr{R}_i$ is a subset of the non-negative real numbers and may be discrete or continuous. The problem is to minimize the end-to-end distortion D in Eq. (2) given a maximum average rate $R = N^{-1}\sum_{i=1}^{N} R_i$.

It results of the mean theorem (i.e., the arithmetic mean of the w_iD_is is not smaller than their geometric mean, with equality if and only if all the terms are equal) that, under the constraint of a given average rate R, the distortion D is minimum if and only if all the w_iD_is are equal, in which case

$$D_{\mathbf{T}}(R) \approx \left(\prod_{i=1}^{N} w_i c_i\right)^{1/N} \left(\prod_{i=1}^{N} \sigma_i^2\right)^{1/N} 2^{-2R}, \tag{4}$$

where σ_i^2 is the variance of Y_i and c_i is the constant associated with the standardized variable of Y_i according to the relation (3). When no transform $\mathbf{T}$ is applied to $\mathbf{X}$, or equivalently when $\mathbf{T}$ is the identity $\mathbf{I}$, the minimum distortion associated with the same maximum average rate R is given by $D_{\mathbf{I}}(R) \approx (\prod_{i=1}^{N} c_i^{\star})^{1/N} (\prod_{i=1}^{N} \sigma_i^{\star 2})^{1/N} 2^{-2R}$, where $\sigma_i^{\star 2}$ is the variance of X_i, and $c_i^{\star}$ is the constant associated with the standardized variable of X_i according to the relation (3).

2.3. Generalized coding gain and maximum reducible bits

In this paragraph we present a criterion called the generalized coding gain (resp. the generalized maximum reducible bits) which is a generalization of the coding gain [3] (resp. the maximum reducible bits) to non-Gaussian signals and non-orthogonal linear transforms.

The distortion rate (4) can be used to define a figure of merit that we call the *generalized coding gain*

$$G^{\star} = \frac{D_{\mathbf{I}}(R)}{D_{\mathbf{T}}(R)} \approx \frac{\left(\prod_{i=1}^{N} c_i^{\star}\right)^{1/N} \left(\prod_{i=1}^{N} \sigma_i^{\star 2}\right)^{1/N}}{\left(\prod_{i=1}^{N} w_i c_i\right)^{1/N} \left(\prod_{i=1}^{N} \sigma_i^2\right)^{1/N}}. \tag{5}$$

It is the factor by which the distortion is reduced because of the linear transform $\mathbf{T}$, assuming high-rate quantization and optimal bit allocation. Taking

the inverse of (4), we obtain the optimal rate distortion: $R_{\mathbf{T}}(D) \approx \frac{1}{N}\sum_{i=1}^{N} h(Y_i) + \frac{1}{2N}\sum_{i=1}^{N} \log_2 w_i - \frac{1}{2}\log_2[12D]$, from which we can define the *generalized maximum reducible bits* $R^\star = R_{\mathbf{I}}(D) - R_{\mathbf{T}}(D) = \frac{1}{2}\log_2 G^\star$. It is the quantity by which the average number of bits to code $\mathbf{X}$ is reduced because of the transform $\mathbf{T}$.

If we assume that $\mathbf{X}$ is Gaussian, then $\mathbf{Y}$ is Gaussian and we have $c_i = c_i^\star = \frac{\pi e}{6}$ $(1 \leqslant i \leqslant N)$, hence the generalized coding gain becomes $G^\star = (\prod_{i=1}^{N} \sigma_i^{\star 2})^{1/N} / (\prod_{i=1}^{N} w_i \sigma_i^2)^{1/N}$. Furthermore, if we suppose that $\mathbf{T}$ is orthogonal, then the w_is are all equal to one and the generalized coding gain is identical to the well-known coding gain G appearing in texts and scholarly books, which is maximized for any Karhunen–Loève basis of $\mathbf{X}$ (see, for example, [3]).

2.4. Criterion satisfied by an optimal transform

Using the following relations (see, e.g., [17] for notions of information theory) $h(\mathbf{X}) = \sum_{i=1}^{N} h(X_i) - I(X_1; \ldots; X_N)$, $h(\mathbf{Y}) = \sum_{i=1}^{N} h(Y_i) - I(Y_1; \ldots; Y_N)$, $h(\mathbf{Y}) = h(\mathbf{X}) + \log_2 |\det \mathbf{T}|$, where $I(X_1; \ldots; X_N)$ and $I(Y_1; \ldots; Y_N)$ denote, respectively, the mutual information between the components of $\mathbf{X}$ and of $\mathbf{Y}$, $h(\mathbf{X})$ and $h(\mathbf{Y})$ the differential entropy of vector $\mathbf{X}$ and $\mathbf{Y}$, respectively, the generalized maximum reducible bits can be expressed as follows:

$$R^\star = \frac{1}{N} I(X_1; \ldots; X_N) - \frac{1}{N} I(Y_1; \ldots; Y_N) - \frac{1}{N} \log_2 |\det \mathbf{T}| - \frac{1}{2N} \log_2 \prod_{i=1}^{N} w_i. \quad (6)$$

Moreover, according to the expression of the w_is, and denoting $\mathrm{Diag}(\mathbf{C})$ for the diagonal matrix having the same main diagonal as $\mathbf{C}$, the last two terms in (6) are equal to $-\frac{1}{2N}\log_2\left(\frac{\det[\mathrm{Diag}(\mathbf{T}^{-\mathrm{T}}\mathbf{T}^{-1})]}{\det[\mathbf{T}^{-\mathrm{T}}\mathbf{T}^{-1}]}\right)$. Hence we deduce the following expression of the generalized maximum reducible bits:

$$R^\star = \frac{1}{N} I(X_1; \ldots; X_N) - \frac{1}{N} I(Y_1; \ldots; Y_N) - \frac{1}{2N} \log_2 \left(\frac{\det[\mathrm{Diag}(\mathbf{T}^{-\mathrm{T}}\mathbf{T}^{-1})]}{\det[\mathbf{T}^{-\mathrm{T}}\mathbf{T}^{-1}]} \right). \quad (7)$$

Let us remark, according to Hadamard's inequality (see e.g., [17]), that for any definite positive matrix $\mathbf{C}$ of order N, the quantity $\log_2\left(\frac{\det[\mathrm{Diag}(\mathbf{C})]}{\det \mathbf{C}}\right)$ is greater or equal than zero, with equality if and only if the matrix $\mathbf{C}$ is diagonal.

It is now clear that the problem of finding a linear transform $\mathbf{T}$ which maximizes the generalized coding gain $G^\star$ defined in (5) is the same problem as finding $\mathbf{T}$ which maximizes $R^\star$, or equivalently, finding the linear transform $\mathbf{T}$ which minimizes the contrast

$$\mathscr{C}(\mathbf{T}) = I(Y_1; \ldots; Y_N) + \frac{1}{2} \log_2 \left(\frac{\det[\mathrm{Diag}(\mathbf{T}^{-\mathrm{T}}\mathbf{T}^{-1})]}{\det[\mathbf{T}^{-\mathrm{T}}\mathbf{T}^{-1}]} \right). \quad (8)$$

The first term of (8) is a measure of the statistical dependence between the transform coefficients Y_i. It is always non-negative, and zero if and only if the variables are statistically independent. As for the second term, it is always non-negative, and zero if and only if the column vectors of $\mathbf{T}^{-1}$ are pairwise orthogonal. Furthermore, if $\mathbf{D}$ is a diagonal matrix, one can verify that $\mathscr{C}(\mathbf{DT}) = \mathscr{C}(\mathbf{T})$, i.e., the contrast is scale invariant. Intuitively, this is not surprising since multiplying each component by a factor λ_i and each quantization step by the same factor has no impact on both the final rate and the end-to-end distortion. We can now state the following theorem.

Theorem 1. *Under the high-rate quantization hypothesis, for entropy-constrained scalar quantizers, mean-square distortion and the condition* $\mathbf{U} = \mathbf{T}^{-1}$, *a linear transform* $\mathbf{T}$ *is optimal in coding if and only if it minimizes the contrast* $\mathscr{C}(\mathbf{T})$ *of relation* (8).

Remark now that the contrast $\mathscr{C}(\mathbf{T})$ is always non-negative, and that it is equal to zero if and only if $\mathbf{T}^{-1}$ is a transform with orthogonal columns which produces independent coefficients.

3. Link between ICA and high-rate transform coding

3.1. Criteria

The criterion (8) may be decomposed into $\mathscr{C}(\mathbf{T}) = \mathscr{C}_{\mathrm{ICA}}(\mathbf{T}) + \mathscr{C}_{\mathrm{O}}(\mathbf{T})$, where $\mathscr{C}_{\mathrm{ICA}}(\mathbf{T}) = I(Y_1; \ldots; Y_N)$ corresponds to the mutual information criterion in ICA, and

$$\mathscr{C}_{\mathrm{O}}(\mathbf{T}) = \frac{1}{2} \log_2 \left[\frac{\det[\mathrm{Diag}(\mathbf{T}^{-\mathrm{T}}\mathbf{T}^{-1})]}{\det[\mathbf{T}^{-\mathrm{T}}\mathbf{T}^{-1}]} \right]. \quad (9)$$

The second term $\mathscr{C}_{\mathrm{O}}(\mathbf{T})$ measures a pseudo-distance to orthogonality of the transform $\mathbf{T}$. In general, the optimal transform $\mathbf{T}_{\mathrm{opt}}$ in transform coding, i.e., the transform which minimizes the contrast (8), will be

different from that $\mathbf{T}_{\mathrm{ICA}}$ which minimizes the first term of (8), i.e., the solution of the ICA problem. It is important to notice here that the classical assumption made in blind source separation problems, that is the observations are obtained from a linear mixing of independent sources, is not really required in the problem of finding the transform that maximizes the generalized coding gain.

The expression of the contrast (8) depends on the definition of the distortion. In this work, we measure the distortion as mean-squared error, therefore it is not surprising that orthogonal transforms are favored over other linear transforms since they are energy-preserving.

3.2. Modified ICA algorithms for coding

In Appendix, we propose two variants of the `icainf` algorithm by Pham [14] for the minimization of the contrast (8). The first one, called *generalized coding gain supremum* (`GCGsup`) gives the transform $\mathbf{T}_{\mathrm{opt}}$ that minimizes the contrast (8), the second, called *orthogonal independent component analysis* (`OrthICA`) finds the orthogonal matrix $\mathbf{T}_{\mathrm{orth}}$ that minimizes the contrast $\mathscr{C}_{\mathrm{ICA}}(\mathbf{T})$.

The minimization of the criterion (8) can be done through a gradient descent algorithm, but a much faster method is the Newton algorithm (which amounts to using the natural gradient [19]). As in [14], because of the multiplicative structure of our optimization problem, we use multiplicative increment of the parameter $\mathbf{T}$ rather than additive increment. Starting with a current estimator $\widehat{\mathbf{T}}$, it consists of expanding $\mathscr{C}(\widehat{\mathbf{T}}+\mathscr{E}\widehat{\mathbf{T}})$ with respect to the matrix $\mathscr{E}$ up to second order and then minimizing the resulting quadratic form in $\mathscr{E}$ to obtain a new estimate. Note that the parameter $\mathscr{E}$ is a matrix of order N. This method requires the computation of the Hessian[2] of $\mathscr{C}(\widehat{\mathbf{T}}+\mathscr{E}\widehat{\mathbf{T}})$ with respect to $\mathscr{E}$, which is quite involved. For this reason, we will approximate it by the Hessian of $\mathscr{C}(\widehat{\mathbf{T}}+\mathscr{E}\widehat{\mathbf{T}})$, computed under the assumption that the transform coefficients Y_i are independent. The method is then referred to as quasi-Newton. Although those simplifications result in a slower convergence speed towards the solution, they cause the robustness of the algorithm to be improved by reducing the risk of divergence when the initial estimator $\widehat{\mathbf{T}}_0$ is far from the final solution. Note that the final solution is the same as that obtained without simplification since the algorithm consists of canceling the first-order terms in the expansion of $\mathscr{C}(\mathbf{T}+\mathscr{E}\mathbf{T})$.

[2]The Hessian of a function of several variables is the matrix of its second partial derivatives.

4. Experimental results

In this section, we are interested in assessing the performances of $\mathbf{T}_{\mathrm{opt}}$ and $\mathbf{T}_{\mathrm{orth}}$ in transform coding. Results included in this paper show coding performances on two synthetic data sets and a natural image data set. The synthetic data sets are used to show that $\mathbf{T}_{\mathrm{opt}}$ and $\mathbf{T}_{\mathrm{orth}}$ can outperform the KLT when used for high-rate transform coding of non-Gaussian signals. As for the second data set, it consists of well-known grayscale natural images which will be used to evaluate the performances of $\mathbf{T}_{\mathrm{opt}}$ and $\mathbf{T}_{\mathrm{orth}}$ in medium-to-low bit-rate image compression.

4.1. Examples of synthetic signals efficiently compressed with `GCGsup` and `OrthICA`.

4.1.1. The tested signals

The first synthetic data set consists of 2^{16} samples of a bidimensional ($N=2$) random vector $\mathbf{X}$ obtained as follows. First, we produce 2^{16} samples of a white vector $\mathbf{S}=[S_1,S_2]^{\mathrm{T}}$ whose ith component is the standardized random variable associated to $S_i'=\mathrm{Sign}(Z_i).|Z_i|^{\alpha}$, where $[Z_1,Z_2]^{\mathrm{T}}$ is a standardized white Gaussian random vector. The exponent α is an arbitrary positive real number. When $\alpha>1$ (resp. $\alpha<1$), S_i is super- (resp. sub-) Gaussian. Then, the vector $\mathbf{X}$ is obtained via a mirror symmetry, according to the operation $\mathbf{X}=\mathbf{BS}$, with $\mathbf{B}=\frac{\sqrt{2}}{2}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}$ an orthogonal matrix. Samples of $\mathbf{S}$ and $\mathbf{X}$ are depicted, respectively, on the first row and the second row in Fig. 2 for four different values of α. The second synthetic data set consists of 2^{16} samples of a bidimensional random vector $\mathbf{X}$ obtained as follows: $\mathbf{X}=\mathbf{BU}$, where $\mathbf{B}$ is given above and $\mathbf{U}=[U_1,U_2]^{\mathrm{T}}$ with U_1 and U_2 independent and identically distributed uniform random variables on $[-1,1]$. Samples of $\mathbf{U}$ and $\mathbf{X}$ are depicted, respectively, in Fig. 3(a) and (b).

4.1.2. Evaluation methodology

Our objective metric to measure the performance of a transform is the generalized coding gain (5). Given both N sources and a transform, the

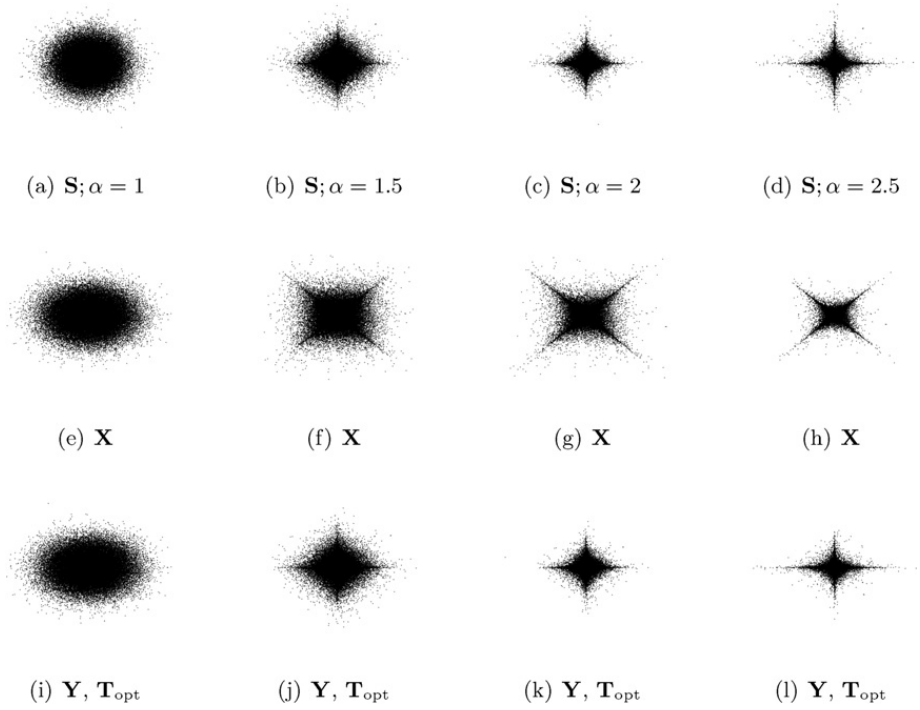


Fig. 2. Samples of **S**, **X** and **Y** (obtained after applying $\mathbf{T}_{\text{opt}}$ on each sample of **X**) for four different values of α.

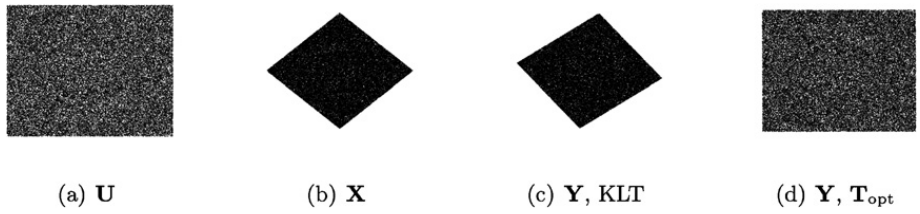


Fig. 3. Samples of **U**, **X** and **Y** (obtained after applying the KLT and $\mathbf{T}_{\text{opt}}$ on **X**).

estimation of (5) requires good estimates of the pdfs of each source as well as each transformed component, which may be very difficult to obtain. In this section, we elaborate on a more "practical" way of evaluating (5) which consists in actually coding the transformed components and measuring both the bit-rate—given here by the first order entropy of the quantized data—and the actual end-to-end distortion.

For each tested signal, the vector $\mathbf{X}$ is first linearly transformed by a transform $\mathbf{T}$ to produce the vector $\mathbf{Y} = [Y_1, \ldots, Y_N]^{\mathrm{T}}$ whose components are coded separately. The ith component Y_i is first high-rate quantized with a uniform scalar quantizer of quantization step q_i. This gives $\widehat{Y}_i$. The bit-rate R_i is then estimated by computing the empirical first-order entropy of $\widehat{Y}_i$ and the inverse transform is applied to $\widehat{\mathbf{Y}} = [\widehat{Y}_1, \ldots, \widehat{Y}_N]^{\mathrm{T}}$ in order to reconstruct an approximation $\widehat{\mathbf{X}} = [\widehat{X}_1, \ldots, \widehat{X}_N]^{\mathrm{T}}$ of $\mathbf{X}$. The distortion is the end-to-end one, $D = \frac{1}{N}\mathrm{E}[\|\mathbf{X} - \widehat{\mathbf{X}}\|^2]$ and the total average rate is the empirical mean of the N rates R_i ($1 \leqslant i \leqslant N$). The optimal allocation of rates between the transform coefficients results in equal weighted distortions $w_i D_i$ (see Section 2). Moreover, using uniform scalar quantizers, bit allocation amounts to choosing a quantization step q_i for each of the N components, and for small q_i the distortion D_i may be well approximated by $q_i^2/12$ [3]. Therefore, the bit allocation follows a simple rule: let c be a constant, then make all the quantization steps $q_1, \ldots, q_N$ such that $w_i D_i = c$. This gives $q_i = \sqrt{12c/w_i}$, for $i = 1, \ldots, N$. When the constant c varies (under the assumption of high-resolution quantization for each component) we obtain the classical asymptotic curve (distortion versus bit-rate, or equivalently bit-rate versus distortion). In our tests, we consider that the hypothesis of high-resolution quantization is valid when for each component Y_i, the relative deviation between the actual distortion $\mathrm{E}[(Y_i - \widehat{Y}_i)^2]$ (where

the expectation is estimated by empirical mean) and $q_i^2/12$ is not greater than 1%. For a given high bit-rate, the ratio between the end-to-end distortion read on the asymptotic curve obtained using the identity transform and that read on the asymptotic curve obtained using $\mathbf{T}$ yields the generalized coding gain of $\mathbf{T}$.

4.1.3. Results

The first set of synthetic data was designed so that the KLT does nothing on it. Indeed, the random vector $\mathbf{X}$ being white, the generalized coding gain $G^\star$ of the KLT is equal to 0 dB. However, the components of $\mathbf{X}$ are not independent, as can be seen on the second row in Fig. 2, and any algorithm among `GCGsup`, `OrthICA` and `icainf` gives the same result: the components Y_i are independent (see third row in Fig. 2), and the generalized coding gain $G^\star$ is the same. The generalized coding gain associated with using the transform returned by `GCGsup` is presented in Table 1 for four different values of α. Except for $\alpha = 1$, i.e., when the components of $\mathbf{X}$ are Gaussian, the generalized coding gain of $\mathbf{T}_{\text{opt}}$ is always greater than zero and becomes more and more important as $|1-\alpha|$ increases, i.e., as the components of $\mathbf{X}$ depart from Gaussianity.

Table 1
Generalized coding gain of $\mathbf{T}_{\text{opt}}$ for the first set of synthetic data

α	1	1.5	2	2.5
$G^\star$ (dB)	0	1.02	2.80	4.78

Table 2
Generalized coding gains of the KLT and $\mathbf{T}_{\text{opt}}$ for the second set of synthetic data

Transform	KLT	$\mathbf{T}_{\text{opt}}$
$G^\star$ (dB)	0.02	1.25

The second set of synthetic data has recently drawn attention of researchers in transform coding. In [6], the authors have investigated the strengths and limitations of the KLT when it is used for transform coding the vector $\mathbf{X} = \mathbf{BU}$. While the KLT for $\mathbf{X}$ is not unique, practical implementations of the KLT give a matrix close to the identity. As a consequence, the diamond distribution of $\mathbf{X}$ remains practically unchanged after applying the KLT on $\mathbf{X}$ (see Fig. 3(c)). Yet $\mathbf{X}$ and $\mathbf{U}$ are not equally good sources for scalar quantization as was already pointed out earlier [21]. This is confirmed in our experiments since any algorithm among `GCGsup`, `OrthICA` and `icainf` transforms the diamond

(a) Lenna (b) Goldhill (c) Mandrill

(d) Peppers (e) Zelda (f) Boat

Fig. 4. Test images.

distribution back into the square distribution of **U** as can be seen in Fig. 3(d). Furthermore, Table 2 shows that the generalized coding gain $G^{\star}$ of the transform returned by `GCGsup` outperforms that obtained with the KLT.

4.2. Application of GCGsup and OrthICA to the compression of natural images

The next experimental tests aim at assessing the performances of `GCGsup` and `OrthICA` when applied to the compression of well-known grayscale natural images (*Lenna*, *Goldhill*, *Mandrill*, *Peppers*, *Zelda* and *Boat*) each of size 512×512 pixels and coded using 8 bits per pixel (bpp) (see Fig. 4). The image coder used in our tests has been designed for experimentation and is not intended to outperform current state-of-the-art image coders such as JPEG2000 [22]. In natural image compression, we are usually interested in performances at medium-to-low bit-rates. In our experiments, we chose to investigate the performances of `GCGsup` and `OrthICA` at bit-rates less or equal than 2 bpp. Note that for bit-rates less than about 1.6 bpp, the transforms returned by `GCGsup` and `OrthICA` can no longer be considered as optimal since the high-resolution hypothesis is no longer valid.[3]

4.2.1. Bases estimation

The modified ICA bases (i.e., the column vectors of **T**) were estimated according to two learning schemes. The first scheme yields 12 different bases (2 per image): for each test image, the algorithms `GCGsup` and `OrthICA` were applied to a training set consisting of 4096 non-overlapping image blocks each of size 8×8 pixels extracted from the test image. The first and second columns of Fig. 5 displays the estimated modified ICA bases as well as the practically achieved KLT bases obtained from the test images *Boat* and *Peppers*, respectively. Also displayed for comparison are the bases obtained using the ICA algorithm `icainf`. As for the second scheme, it yields only two different bases. The modified ICA bases were learned from one training set consisting of 12 288 non-overlapping image blocks each of size 8×8 pixels extracted from three test images (*Lenna*, *Goldhill* and *Boat*). The third column of Fig. 5 displays the estimated modified ICA bases as well as the KLT basis (denoted KLT*). The transform whose column vectors were estimated using the algorithm `OrthICA` (resp. `GCGsup`) is denoted $\mathbf{T}^{\star}_{\mathrm{orth}}$ (resp. $\mathbf{T}^{\star}_{\mathrm{opt}}$). The ICA basis obtained with the `icainf` algorithm for this training set is also displayed for comparison and denoted $\mathbf{T}^{\star}_{\mathrm{ICA}}$.

Examining Fig. 5 closely reveals that the features found with `GCGsup` and `OrthICA` are much more localized in space than the checkerboard-like basis vectors obtained with the KLT. Note also the more pronounced edge-like nature of the modified ICA bases, regardless of the learning scheme employed. Other similar experiments with ICA [23] have produced comparable results. The computation of the pseudo-distance to orthogonality (see Eq. (9)) reveals that, unlike the bases obtained with the ICA algorithm `icainf`, those estimated with `GCGsup` are quasi-orthogonal as can be seen in Table 3. As mentioned earlier, it is not surprising that orthogonal transforms are favored over other linear transforms since we measure the distortion as mean-squared error and orthogonal transforms are energy-preserving.

4.2.2. Medium-to-high bit-rates

Table 4 shows estimations of the generalized coding gain for each tested transform and each test image. The average generalized coding gain computed over the set of test images is also given. The estimation method used is the same as that described in Section 4.1.2. Looking at the average values of the generalized coding gain reveals that, whatever the learning scheme, the modified ICA transforms perform best followed, respectively, by the 2-D DCT, the KLT and, far behind, the ICA transform. The coding gain of any of the modified ICA transforms relative to the 2-D DCT is about 0.3 dB (resp. 0.1 dB) when the first (resp. second) learning scheme is used suggesting that a transform-based image coder could benefit from using any of the modified ICA transforms. However, these results should be taken with care since they are meaningful only under the high-resolution hypothesis (i.e., in general, at medium-to-high bit-rates). Yet, in image compression, we are most interested in performances at medium-to-low bit-rates, where the high-resolution hypothesis is no longer valid. Additional precaution should be taken for the bases estimated according to the first learning scheme. In this case, the basis vectors need to be transmitted

[3] As in Section 4.1.2, we consider that the hypothesis of high-resolution quantization is valid when for each component Y_i, the relative deviation between the actual distortion $\mathrm{E}[(Y_i - \widehat{Y}_i)^2]$ (where the expectation is estimated by empirical mean) and $q_i^2/12$ is not greater than 1%.

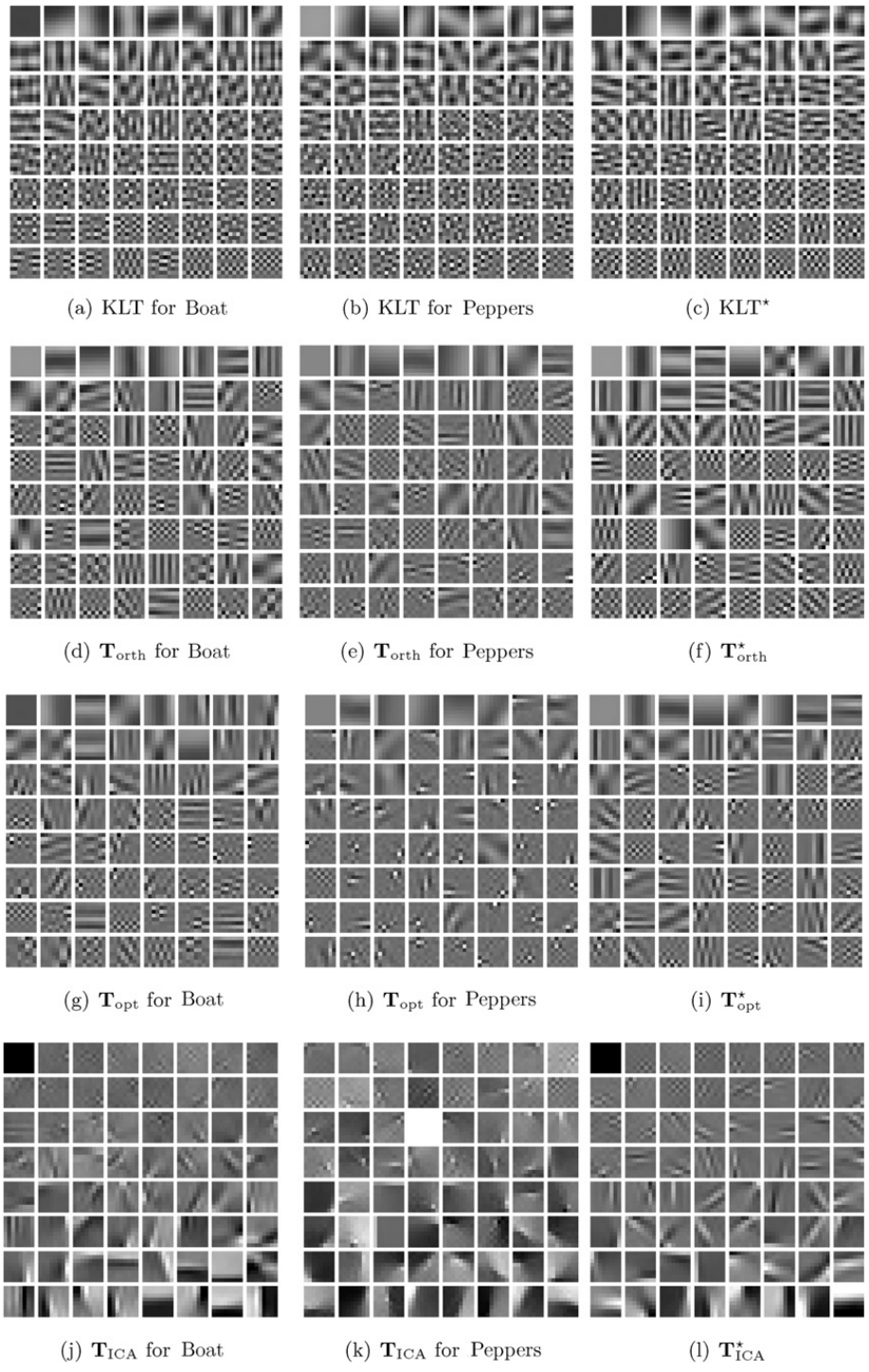


Fig. 5. KLT, $\mathbf{T}_{\mathrm{orth}}$, $\mathbf{T}_{\mathrm{opt}}$ and $\mathbf{T}_{\mathrm{ICA}}$ basis vectors obtained from boat and peppers (on first and second column, respectively) and KLT*, $\mathbf{T}^{\star}_{\mathrm{orth}}$, $\mathbf{T}^{\star}_{\mathrm{opt}}$ and $\mathbf{T}^{\star}_{\mathrm{ICA}}$ basis vectors obtained from the image class training set.

Table 3
Distance to orthogonality (in bpp) of $\mathbf{T}_{\mathrm{opt}}$, $\mathbf{T}^{\star}_{\mathrm{opt}}$, $\mathbf{T}_{\mathrm{ICA}}$ and $\mathbf{T}^{\star}_{\mathrm{ICA}}$ for each test image

	Lenna	Goldhill	Mandrill	Peppers	Zelda	Boat
$\mathbf{T}_{\mathrm{opt}}$	0.009	0.008	0.008	0.016	0.004	0.012
$\mathbf{T}^{\star}_{\mathrm{opt}}$	0.006	0.006	0.006	0.006	0.006	0.006
$\mathbf{T}_{\mathrm{ICA}}$	1.955	0.711	0.356	1.108	2.435	0.765
$\mathbf{T}^{\star}_{\mathrm{ICA}}$	0.305	0.305	0.305	0.305	0.305	0.305

Table 4
Generalized coding gain (in dB) of the KLT, $\mathbf{T}_{\text{orth}}$, $\mathbf{T}_{\text{opt}}$, KLT*, $\mathbf{T}^{\star}_{\text{orth}}$, $\mathbf{T}^{\star}_{\text{opt}}$, $\mathbf{T}_{\text{ICA}}$, $\mathbf{T}^{\star}_{\text{ICA}}$ and 2-D DCT for each test image

	Lenna	Goldhill	Mandrill	Peppers	Zelda	Boat	Average
KLT	18.13	15.22	6.99	17.23	19.91	15.35	15.47
$\mathbf{T}_{\text{orth}}$	18.58	15.62	7.42	17.89	20.12	15.96	15.93
$\mathbf{T}_{\text{opt}}$	18.60	15.68	7.49	17.84	20.12	15.94	15.94
KLT*	17.76	15.09	6.98	17.08	19.26	14.79	15.16
$\mathbf{T}^{\star}_{\text{orth}}$	18.34	15.40	7.27	17.63	19.89	15.74	15.73
$\mathbf{T}^{\star}_{\text{opt}}$	18.31	15.46	7.28	17.70	19.84	15.72	15.71
$\mathbf{T}_{\text{ICA}}$	6.88	11.63	5.52	11.47	5.36	11.62	8.74
$\mathbf{T}^{\star}_{\text{ICA}}$	16.50	13.75	5.51	16.03	17.83	13.95	13.92
2-D DCT	18.25	15.37	7.17	17.50	19.87	15.61	15.62

Last column yields the average generalized coding gain of each transform computed over the set of test images.

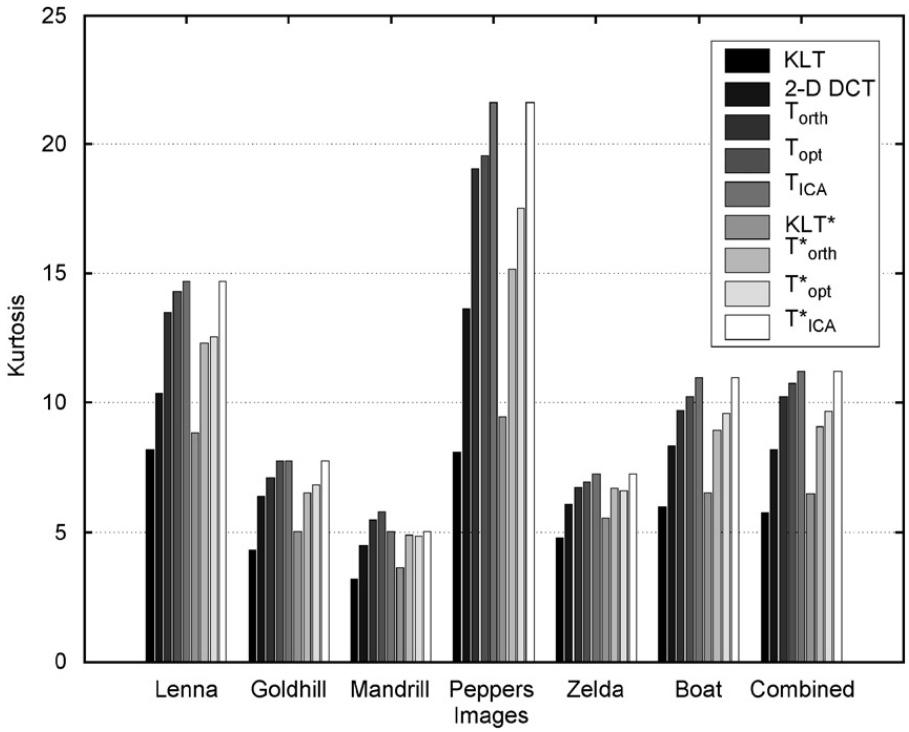


Fig. 6. Average kurtosis computed over 63 transformed components (the component which is equivalent to the DC component of the KLT was omitted).

(coded) with the image resulting in an extra coding cost which has not been incorporated in the method used for estimating the generalized coding gain.

Another motivation for using the modified ICA bases in compression lies in the transform coefficients distribution which tends to be heavy-tailed, more suited for quantization and entropy coding than the transform coefficients of the KLT and 2-D DCT, which tend to a normal distribution. Fig. 6 displays for each transform and each image the average kurtosis computed over 63 transformed components (the component which is equivalent to the DC component of the KLT was omitted). The kurtosis was normalized so that it is equal to zero in the case of Gaussian samples. The average of this *average kurtosis* was computed over the set of test images. The result for each transform is displayed over the label "combined" in Fig. 6. Results show that the kurtosis obtained with the modified ICA bases are greater than those obtained with the KLT and 2-D DCT, regardless of the transform and image considered. Note that the ICA bases exhibit

Table 5
PSNR (dB) versus target bit-rate for each test image and each transform

Image	Target bit-rate (bpp)	KLT	$\mathbf{T}_{\text{opt}}$	$\mathbf{T}_{\text{orth}}$	KLT*	2-D DCT	$\mathbf{T}^\star_{\text{opt}}$	$\mathbf{T}^\star_{\text{orth}}$
Lenna	2	42.37	42.72	42.76	42.92	43.21	43.30	**43.40**
	1	37.32	37.77	37.82	37.95	38.44	38.54	**38.55**
	0.5	32.56	32.86	33.05	34.31	34.83	34.95	**35.00**
	0.25	26.15	25.34	26.40	30.72	31.18	31.30	**31.32**
Goldhill	2	39.45	39.77	39.82	40.17	**40.80**	40.50	40.48
	1	34.06	34.29	34.34	34.93	**35.41**	35.24	35.22
	0.5	30.03	30.13	30.34	31.50	**31.92**	31.85	31.81
	0.25	25.57	24.76	25.66	28.79	**29.20**	29.15	29.09
Mandrill	2	32.23	32.79	32.69	33.20	**33.55**	33.48	33.43
	1	26.55	27.03	27.02	27.56	**27.87**	27.81	27.81
	0.5	23.02	23.21	23.22	24.31	**24.49**	24.48	24.48
	0.25	20.29	20.37	20.35	22.14	**22.34**	22.30	22.29
Peppers	2	40.43	40.86	40.90	41.13	41.49	**41.70**	41.68
	1	35.71	36.11	36.15	36.36	36.71	**36.84**	36.79
	0.5	31.85	32.23	32.27	33.44	33.91	**34.09**	33.99
	0.25	25.53	24.24	24.18	30.18	30.74	**30.88**	30.87
Zelda	2	45.21	45.41	45.39	45.13	45.85	45.86	**45.96**
	1	40.16	40.30	40.33	40.61	40.94	**41.02**	41.01
	0.5	35.97	36.31	36.28	37.47	37.75	**37.93**	37.90
	0.25	29.09	29.14	29.17	34.03	34.49	34.57	**34.60**
Boat	2	41.67	42.15	42.25	42.04	**43.27**	42.75	42.87
	1	35.39	35.85	36.09	36.25	**37.51**	37.14	37.13
	0.5	30.16	30.36	30.65	31.91	**32.88**	32.73	32.62
	0.25	24.37	23.88	24.56	28.29	**29.23**	29.09	29.01

The bold value is used to highlight the transform that performs best on a given image and at a given target bit-rate.

even higher kurtosis, yet yielding the worst generalized coding gains among all tested transforms. The explanation for this is straightforward: ICA bases were estimated using the ICA algorithm `icainf` which aims at minimizing the mutual information between the components, i.e., only the first term in Eq. (8). This is done without putting any orthogonality constraint on the basis vectors resulting in a transform which is far from being orthogonal (see Table 3). This result highlights the importance of trading-off between independence and orthogonality—when the distortion is measured using mean-squared error. This is well illustrated by Eq. (8) (a discussion about orthogonality and independence can also be found in [4]).

4.2.3. Medium-to-low bit-rates

The image coder used in our experiment is a transform coder originally developed by Davis.[4] It is very modular and allows for simple replacements of individual components (quantizer, entropy coder, transform). It was modified so that it resembles a JPEG-like coder. The image to be coded is first "tiled" into blocks of 8×8 pixels each, then each tile is represented into a new basis using one of the tested transforms. The bases obtained using the first learning scheme are transmitted with the image since they are data-dependent bases. As for the bases estimated using the second learning scheme, they are *not* transmitted with the image. Quantization steps are chosen to minimize the end-to-end distortion (2) subject to bit-rate constraint. The bit allocation procedure is based on integer programming algorithms described in [24] which provide optimal or near-optimal allocations for the quantizers included here. Entropy coding of the quantizer output is carried out by an adaptive arithmetic coder.

4.2.4. Compression results

We now compare the compression performances of the previously mentioned transforms: KLT, $\mathbf{T}_{\text{opt}}$, $\mathbf{T}_{\text{orth}}$, KLT*, $\mathbf{T}^\star_{\text{opt}}$, $\mathbf{T}^\star_{\text{orth}}$ and 2-D DCT. The results are displayed in Table 5, in which we present the

[4] http://www.geoffdavis.net/dartmouth/wavelet/wavelet.html.

peak signal-to-noise ratio (PSNR) as a function of bit-rate for each test image. Whatever the image, the following characteristics can be observed:

(1) The transform codes based on KLT*, $\mathbf{T}^\star_{\text{opt}}$, $\mathbf{T}^\star_{\text{orth}}$ and the 2-D DCT perform better than those based on the KLT, $\mathbf{T}_{\text{opt}}$ and $\mathbf{T}_{\text{orth}}$, regardless of the bit-rate. The poor coding performances of the KLT, $\mathbf{T}_{\text{opt}}$ and $\mathbf{T}_{\text{orth}}$ are mainly due to the coding penalty resulting from coding the basis vectors (11 bits were allocated on average to each matrix coefficient resulting in a coding precision of 10^{-3}). In our tests, we observed that the coding penalty is about 1 dB at 2 bpp and 1 bpp, about 2 dB at 0.5 bpp and can reach up to 5 dB at 0.25 bpp. The rate–distortion pairs obtained with the KLT, $\mathbf{T}_{\text{opt}}$ and $\mathbf{T}_{\text{orth}}$ without coding the basis vectors (not displayed here for the sake of clarity) were close to those obtained with, respectively, KLT*, $\mathbf{T}^\star_{\text{opt}}$ and $\mathbf{T}^\star_{\text{orth}}$ (no meaningful performance difference could be observed).

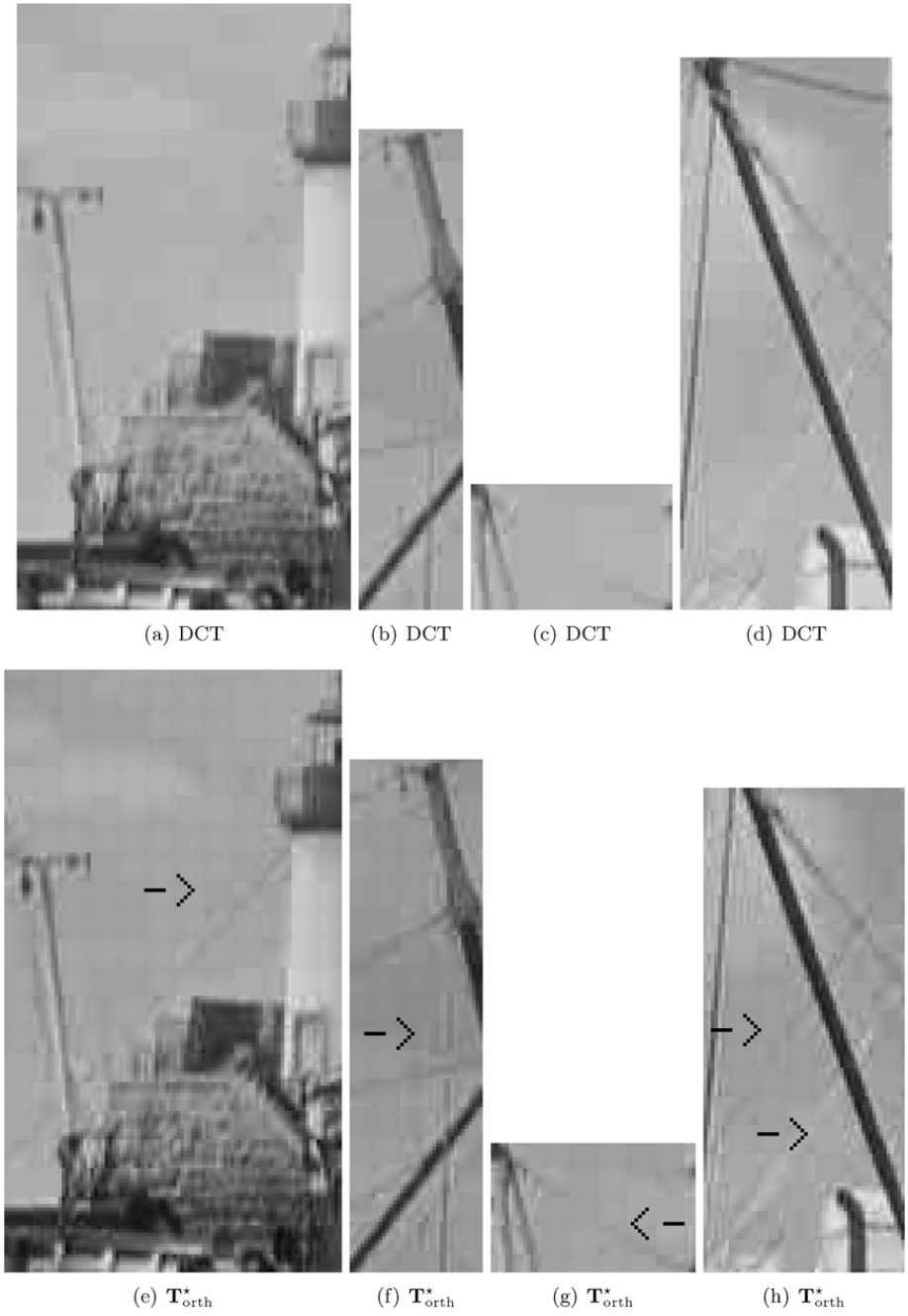


Fig. 7. Zoom on parts of image *Boat* coded at about 0.5 bpp. Black arrows point towards details which are not present or blurred on the corresponding image coded with the 2-D DCT.

(2) The PSNR at 1 bpp is about 5−6 dB lower than that at 2 bpp. This is consistent with our observation that the high-resolution hypothesis is verified for bit-rates greater than about 1.6 bpp (above this rate, the PSNR increases about ~6 dB per bits).

(3) Whatever the bit-rate, no meaningful performance difference can be observed between the class-adapted transform codes based on $\mathbf{T}^{\star}_{\text{opt}}$ and those based on $\mathbf{T}^{\star}_{\text{orth}}$. For medium-to-high bit-rates (i.e., greater than about 1.6 bpp), this result can be more or less predicted by looking at the values of the corresponding generalized coding gains (see Table 4). For bit-rates greater than 1 bpp, the performance difference between either $\mathbf{T}^{\star}_{\text{opt}}$ or $\mathbf{T}^{\star}_{\text{orth}}$ and KLT* is equal to about the difference between their corresponding generalized coding gains suggesting that the performance of our entropy coder is close to that of a perfect first-order entropy coder. In the low bit-rate region (<1 bpp), the performance difference tends to become smaller. As for the 2-D DCT, its performance is comparable with that of any of the class-adapted modified ICA bases. Thus, our approach has made it possible to learn two bases which are competitive with the well-known 2-D DCT basis according to the standard PSNR measure.

Visual inspection of the image quality was also carried out. Although the KLT*-coded images have worse PSNR than the others, no obvious difference in terms of visual quality could be seen. However, when looking further into the details of the reconstructed image *Boat* in Fig. 7 (the tested transform is $\mathbf{T}^{\star}_{\text{orth}}$ and the target bit-rate was set equal to 0.5 bpp), some meaningful visual quality difference can be seen. Black arrows point towards details which are not present or blurred on the corresponding image coded with the 2-D DCT. These details represent lines (e.g., some ropes) which are well preserved with $\mathbf{T}^{\star}_{\text{orth}}$. These results suggest that the class-adapted modified ICA bases are better suited to coding fine details such as lines and edges compared to the 2-D DCT. This is not quite surprising given the more pronounced edge-like nature of the modified ICA bases (see Fig. 5) (Table 5).

5. Conclusion

This paper addresses the problem of finding optimal 1-D linear transforms in transform coding without the classical assumption of Gaussianity. This paper emphasizes a new point of view in variable-rate transform coding by showing, under the high-resolution hypothesis, that the problem of finding the optimal 1-D linear transform may be recast as a modified independent component analysis (ICA) problem. Two new modified ICA algorithms, called `GCGsup` and `OrthICA`, are introduced for computing the optimal 1-D linear transform and the optimal 1-D orthogonal transform, respectively.

Experimental results included in this paper show coding performances of `GCGsup` and `OrthICA` on two synthetic data sets and a natural image data set. Experiments carried out with the synthetic data sets show that the new transforms can outperform the KLT when used for high-rate transform coding of non-Gaussian signals. When applied to the compression of some well-known natural images, `GCGsup` and `OrthICA` have proved (1) to be comparable to the classical 2-D DCT according to the PSNR measure and (2) to yield better visual image quality (better preservation of lines and edges) than the 2-D DCT.

Acknowledgments

The authors are grateful to Pierre Duhamel for his very helpful comments as well as Jacques Weidig and Jean-Louis Gutzwiller who took part in the simulation work.

Appendix A

A.1. *Algorithm* `GCGsup`

In order to simplify some expressions, we use the Neperian logarithm instead of the base 2 logarithm. Given that $I(Y_1,\ldots,Y_N)=\sum_i h(Y_i)-h(\mathbf{Y})$ and $h(\mathbf{Y})=h(\mathbf{X})+\log|\det\mathbf{T}|$, and since the term $h(\mathbf{X})$ does not depend on $\mathbf{T}$, minimizing the contrast (8) is the same as minimizing $\widetilde{\mathscr{C}}(\mathbf{T})=\mathscr{C}_{\text{O}}(\mathbf{T})+\widetilde{\mathscr{C}}_{\text{ICA}}(\mathbf{T})$ where $\widetilde{\mathscr{C}}_{\text{ICA}}(\mathbf{T})=\sum_{i=1}^{N}h(Y_i)-\log|\det\mathbf{T}|$. Using the results of [20] it can be seen that the Taylor expansion of $\widetilde{\mathscr{C}}_{\text{ICA}}(\mathbf{T}+\mathscr{E}\mathbf{T})$ up to second order may be approximated as follows:

$$\begin{aligned}\widetilde{\mathscr{C}}_{\text{ICA}}(\mathbf{T}+\mathscr{E}\mathbf{T}) &= \widetilde{\mathscr{C}}_{\text{ICA}}(\mathbf{T})+\sum_{1\leqslant i\neq j\leqslant N}\mathrm{E}[\psi_{Y_i}(Y_i)Y_j]\mathscr{E}_{ij}\\ &\quad+\frac{1}{2}\sum_{1\leqslant i\neq j\leqslant N}\{\mathrm{E}[\psi^2_{Y_i}(Y_i)]\mathrm{E}[Y_j^2]\mathscr{E}^2_{ij}+\mathscr{E}_{ij}\mathscr{E}_{ji}\}+\cdots,\end{aligned} \tag{10}$$

where the function ψ_{Y_i} is equal to the derivative of $-\log p(y_i)$—$p(y_i)$ denoting the Y_i pdf—and is known as the score function. This approximation concerns only the second-order terms in the expansion, but *not the first-order terms*. It relies essentially on the assumption of independent transform coefficients, which may not be valid if the solution of the ICA problem is far from the solution that minimizes the contrast (8). But it is quite useful since it leads to a decoupling in the quadratic form of the expansion. Let $\mathbf{M}=\mathbf{T}^{-\mathrm{T}}\mathbf{T}^{-1}$. One may verify that the Taylor expansion of $\mathscr{C}_{\mathrm{O}}(\mathbf{T}+\mathscr{E}\mathbf{T})$ with respect to $\mathscr{E}$ and around $\mathscr{E}=\mathbf{0}$, up to second order, is given by $\mathscr{C}_{\mathrm{O}}(\mathbf{T}+\mathscr{E}\mathbf{T})=\mathscr{C}_{\mathrm{O}}(\mathbf{T})-\sum_{1\leqslant i\neq j\leqslant N}\frac{M_{ji}}{M_{ii}}\mathscr{E}_{ji}-\frac{1}{2}\sum_{1\leqslant i\neq j\leqslant N}[\mathscr{E}_{ij}\mathscr{E}_{ji}-2\frac{M_{ji}}{M_{jj}}\mathscr{E}_{ii}\mathscr{E}_{ji}]+\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\sum_{k=1,k\neq i}^{N}[(\frac{M_{jk}}{2M_{ii}}-\frac{M_{ij}M_{ik}}{M_{ii}^2})\mathscr{E}_{ji}\mathscr{E}_{ki}+\frac{M_{kj}}{M_{kk}}\mathscr{E}_{ji}\mathscr{E}_{ik}]+\cdots$. The quadratic form associated with the above expansion is quite involved and is not positive. One possible approximation consists in neglecting the non-diagonal elements of $\mathbf{M}$, which amounts to assuming that the optimal linear transform is close to an orthogonal transform. Under this hypothesis, one may verify that

$$\mathscr{C}_{\mathrm{O}}(\mathbf{T}+\mathscr{E}\mathbf{T})\approx\mathscr{C}_{\mathrm{O}}(\mathbf{T})-\sum_{1\leqslant i\neq j\leqslant N}\frac{M_{ji}}{M_{ii}}\mathscr{E}_{ji}+\frac{1}{2}\sum_{1\leqslant i\neq j\leqslant N}\left[\frac{M_{jj}}{M_{ii}}\mathscr{E}_{ji}^2+\mathscr{E}_{ji}\mathscr{E}_{ij}\right]+\cdots. \tag{11}$$

The quadratic form associated with the above expansion is now positive, but not positive definite. However, this is sufficient for the matrix associated with the quadratic form of the Taylor expansion of $\widetilde{\mathscr{C}}(\mathbf{T})$ to be positive definite, which ensures the stability of the iterative algorithm. Finally, by adding Eq. (10) and approximation (11) we obtain an approximation of $\widetilde{\mathscr{C}}(\mathbf{T}+\mathscr{E}\mathbf{T})$ up to second order of $\mathscr{E}_{ij}$ and the iteration consists explicitly of solving the linear equations

$$\begin{bmatrix}\mathrm{E}[\psi_{Y_i}^2(Y_i)]\mathrm{E}[Y_j^2]+\dfrac{M_{ii}}{M_{jj}} & 2\\ 2 & \mathrm{E}[\psi_{Y_j}^2(Y_j)]\mathrm{E}[Y_i^2]+\dfrac{M_{jj}}{M_{ii}}\end{bmatrix}\times\begin{bmatrix}\mathscr{E}_{ij}\\ \mathscr{E}_{ji}\end{bmatrix}=\begin{bmatrix}\dfrac{M_{ij}}{M_{jj}}-\mathrm{E}[\psi_{Y_i}(Y_i)Y_j]\\ \dfrac{M_{ji}}{M_{ii}}-\mathrm{E}[\psi_{Y_j}(Y_j)Y_i]\end{bmatrix}.$$

The indeterminate diagonal terms $\mathscr{E}_{ii}$ are arbitrarily fixed to zero. Then the estimator $\widehat{\mathbf{T}}$ is left multiplied by $\mathbf{I}+\mathscr{E}$ in order to update it. In this expression, the pdfs being unknown, the score function $\psi_{Y_i}(y_i)$ is replaced by an estimation (see [14]) and the expectations are estimated by empirical means.

A.2. *Algorithm* `OrthICA`

In this section, we propose to find the orthogonal transform that minimizes the contrast (8). Since the second term of (8) vanishes for any orthogonal matrix $\mathbf{T}$, this amounts to finding the orthogonal transform which minimizes the first term of (8), or equivalently, which minimizes $\widetilde{\mathscr{C}}_{\mathrm{ICA}}(\mathbf{T})$. If the matrix $\mathbf{T}$ is orthogonal, so is $\mathbf{T}+\mathscr{E}\mathbf{T}$, providing that $\mathbf{I}+\mathscr{E}$ be orthogonal. This last condition will be satisfied up to second order if $\mathscr{E}$ is anti-symmetric, since $(\mathbf{I}+\mathscr{E})^{\mathrm{T}}(\mathbf{I}+\mathscr{E})=\mathbf{I}+\mathscr{E}^{\mathrm{T}}\mathscr{E}$ differs from the identity only by second-order terms. Let $\mathscr{E}$ be anti-symmetric. The Taylor expansion of $\widetilde{\mathscr{C}}_{\mathrm{ICA}}(\mathbf{T}+\mathscr{E}\mathbf{T})$ becomes $\widetilde{\mathscr{C}}_{\mathrm{ICA}}(\mathbf{T}+\mathscr{E}\mathbf{T})=\widetilde{\mathscr{C}}_{\mathrm{ICA}}(\mathbf{T})+\sum_{1\leqslant i<j\leqslant N}\{\mathrm{E}[\psi_{Y_i}(Y_i)Y_j]-\mathrm{E}[\psi_{Y_j}(Y_j)Y_i]\}\mathscr{E}_{ij}+\frac{1}{2}\sum_{1\leqslant i<j\leqslant N}\mathscr{E}_{ij}^2[\mathrm{E}[\psi_{S_i}^2(Y_i)]\mathrm{E}[Y_j^2]+\mathrm{E}[\psi_{Y_j}^2(Y_j)]\mathrm{E}[Y_i^2]-2]+\cdots$, and the minimization of the second term in the above expansion yields

$$\mathscr{E}_{ij}=\frac{\mathrm{E}[\psi_{Y_j}(Y_j)Y_i]-\mathrm{E}[\psi_{Y_i}(Y_i)Y_j]}{\mathrm{E}[\psi_{Y_i}^2(Y_i)]\mathrm{E}[Y_j^2]+\mathrm{E}[\psi_{Y_j}^2(Y_j)]\mathrm{E}[Y_i^2]-2}. \tag{12}$$

Actually, $\mathbf{T}+\mathscr{E}\mathbf{T}$ is not a true orthogonal transform. This may be overcome by replacing $\mathbf{T}+\mathscr{E}\mathbf{T}$ with $e^{\mathscr{E}}\mathbf{T}=(\mathbf{I}+\mathscr{E}+\mathscr{E}^2/2!+\cdots)\mathbf{T}$, which is an orthogonal matrix differing from $\mathbf{T}+\mathscr{E}\mathbf{T}$ only by second-order terms.

References

[1] G. Wallace, Overview of JPEG (ISO/CCITT) still image compression standard, Commun. ACM 4 (4) (1991) 30–40.
[2] V.K. Goyal, J. Zhuang, M. Vetterli, Transform coding with backward adaptive updates, IEEE Trans. Inform. Theory 46 (July 2000) 1623–1633.
[3] A. Gersho, R. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publisher, Dordrecht, 1992.
[4] V.K. Goyal, Theoretical foundations of transform coding, IEEE Signal Process. Mag. 18 (5) (2001) 9–21.
[5] J.-Y. Huang, P.M. Schultheiss, Block quantization of correlated Gaussian random variables, IEEE Trans. Commun. COM-11 (September 1963) 289–296.

[6] M. Effros, H. Feng, K. Zeger, Suboptimality of the Karhunen–Loève transform for transform coding, IEEE Trans. Inform. Theory 50 (8) (2004) 1605–1619.
[7] S. Mallat, F. Falzon, Analysis of low bit rate image transform coding, IEEE Trans. Signal Process. 46 (4) (1998) 1027–1042.
[8] S. Jana, P. Moulin, Optimality of KLT for high-rate transform coding of Gaussian vector-scale mixtures: application to reconstruction, estimation and classification, IEEE Trans. Inform Theory 52 (9) (2006) 4049–4067.
[9] D. Mary, D. Slock, A theoretical high-rate analysis of causal versus unitary online transform coding, IEEE Trans. Signal Process. 54 (4) (2006) 1472–1482.
[10] A. Bell, T. Sejnowski, The 'independent components' of natural scenes are edge filters, Vision Res. 37 (1997) 3327–3338.
[11] A. T. Puga, A. P. Alves, An experiment on comparing PCA and ICA in classical transform image coding, in: Proceedings of the 1st Workshop on Blind Seperation and ICA, 1998, pp. 105–108.
[12] S. Marusic, G. Deng, ICA-FIR based image redundancy reduction, in: Proceedings of the 1st International Workshop on ICA and Signal Separation, Aussois, France, 1999, pp. 191–196.
[13] A. Ferreira, M. Figueiredo, Class-adapted image compression using independent component analysis, in: Proceedings of the IEEE International Conference on Image Processing—ICIP'2003, Barcelona, Spain, 2003.
[14] D.T. Pham, Fast algorithms for mutual information based independent component analysis, IEEE Trans. Signal Process. 52 (10) (2004) 2690–2700.
[15] M. Narozny, M. Barret, D.T. Pham, I.P. Akam Bita, Modified ICA algorithms for finding optimal transforms in transform coding, in: Proceedings of the IEEE 4th International Symposium on Image and Signal Processing and Analysis, Zagreb, Croatia, 2005, pp. 111–116.
[16] M. Narozny, M. Barret, ICA-based algorithms applied to image coding, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Hawaii, USA, April 2007.
[17] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.
[18] I. Witten, R. Neal, J. Cleary, Arithmetic coding for data compression, Comm. ACM 30 (6) (1987) 519–540.
[19] S. Amari, Natural gradient works efficiently in learning, Neural Comput 10 (2) (1998) 251–276.
[20] D.T. Pham, Entropy of a variable slightly contaminated with another, IEEE Signal Process. Lett. 12 (7) (2005) 536–539.
[21] H. Feng, M. Effros, On the rate–distortion performance and computational efficiency of the Karhunen–Loève transform for lossy data compression, IEEE Trans. Image Process. 11 (2) (2002) 113–122.
[22] M. W. Marcellin, M. Gormish, A. Bilgin, M. Boliek, An overview of JPEG2000, in: Proceedings of Data Compression Conference, Snowbird, Utah, March 2000.
[23] A. Hyvärinen, J. Tarhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.
[24] Y. Shoham, A. Gersho, Efficient bit allocation for an arbitrary set of quantizers, IEEE Trans. Acoustics, Speech Signal Process. 36 (9) (1988) 1445–1453.

Lossy Hyperspectral Images Coding with Exogenous Quasi Optimal Transforms

Michel Barret, Jean-Louis Gutzwiller
SUPELEC, IMS Team
2 rue É. Belin
57070 Metz, France
e-mail: michel.barret@supelec.fr
jean-louis.gutzwiller@supelec.fr

Isidore Paul Akam Bita, Florio Dalla Vedova
LUXSPACE Sarl,
Chateau de Betzdorf,
L-6815, Betzdorf, Luxembourg
e-mail: akambita@luxspace.lu
fdallavedova@luxspace.lu

Abstract

It is well known in transform coding that the Karhunen-Loève Transform (KLT) can be suboptimal for non Gaussian sources. However in many applications using JPEG2000 Part 2 codecs, the KLT is generally considered as the optimal linear transform for reducing redundancies between components of hyperspectral images. In previous works, optimal spectral transforms (OST) compatible with the JPEG2000 Part 2 standard have been introduced, performing better than the KLT but with an heavier computational cost. In this paper, we show that the OST computed on a learning basis constituted of Hyperion hyperspectral images issued from one sensor performs very well, and even better than the KLT, on other images issued from the same sensor.

1. Introduction

These last years, research activities on hyperspectral images compression have been strengthened, due to the development of sensors that supply larger and larger amount of data. The end-users of such images become also more numerous and miscellaneous, therefore there is an interest of designing codecs of hyperspectral images that are not application-dependent. The JPEG2000 Part2 standard offers such a possibility with two variants for reducing spectral redundancies between components of the encoded image: by applying either a 1-D discret wavelet transform (DWT), or a 1-D linear transform. Each component is then split in subbands with a 2-D DWT before scalar quantizations and an entropy coding with Taubman's EBCOT algorithm [1], [2]. Moreover, although the rate

This work wa partially supported by ESA through an Innovation Triangle Initiative (ITI).

1068-0314/09 $25.00 © 2009 IEEE
DOI 10.1109/DCC.2009.8

allocation between image components is outside the scope of the JPEG2000 standard, it is a major point in the design of JPEG2000 encoder. In [3], the authors show that the EBCOT post-compression rate-distortion (PCRD) optimizer performing simultaneously across all the codeblocks from the entire image give the best rate allocation between components. They used a 1-D DWT for reducing spectral redundancies, however the result remains valid with a linear transform. When the spectral decorrelation is achieved via the Karhunen-Loève Transform (KLT) or a Principal Component Analysis (PCA) instead of a 1-D DWT, the performances are significantly improved at low, medium and high bit-rates ([4], [5]), even when the DWT adapts to the encoded image [6], [7]. The main drawback of the data-dependent KLT (or PCA) is its heavy computational cost (due to the calculation of a covariance matrix across the data) therefore low-complexity computations of the covariance matrix, based on the law of large numbers, have been proposed to reduce the computational burden [5], [8].

Another strategy exists to reduce the complexity of KLT based codecs. It consists of computing the KLT on a set of images (called the learning basis) issued from one (and only one) sensor in order to obtain an efficient, although sub-optimal, spectral transform that can be applied to any image issued from the same sensor. This strategy has been successfully applied in [7] for on-board compression of multispectral images. The KLT computed on the learning basis is called exogenous KLT when it is applied on other images. In the present paper, we use the same strategy on hyperspectral images: an exogenous optimal spectral transform is computed from a learning basis then it is applied on other images. The major difference with the approach presented in [7] is that in our tests we shall use instead of the KLT the orthogonal optimal spectral transform (OST) introduced in [9]. Further, we compare the performances in coding of the exogenous OST, the exogenous KLT and also the actual KLT and OST, with different measures of distortion : Signal to Noise Ratio (SNR), Maximum Absolute Difference (MAD), Mean Absolute Error (MAE) and Maximum Spectral Angle (MSA). Indeed, it is well-known that providing the mean square error as the only estimate of distortion is not sufficient to assess the quality of a codec for hyperspectral images [10]. The hyperspectral images used in our tests are issued from an Hyperion imaging spectrometer of the NASA[1]. Before introducing the quasi optimal transforms used in our tests, we recall some well known results in coding lessons.

In a conventional transform coder [11], [12], an input vector $\mathbf{X}$ is first transformed into another vector $\mathbf{Y} = \mathbf{T}\mathbf{X}$ of the same dimension. The components of that vector are then described to the decoder using independent scalar quantizers on the coefficients. Finally, the decoder reconstructs the quantized transform vector $\widehat{\mathbf{Y}}$ and then uses a linear transformation $\mathbf{U}$ to get an estimate of the original input vector $\widehat{\mathbf{X}} = \mathbf{U}\widehat{\mathbf{Y}}$. If the mean-square distortion : $\mathrm{E}[\|\mathbf{X} - \widehat{\mathbf{X}}\|^2]$ is considered, where E denotes mathematical expectation, and if the input vector is Gaussian, then the optimal transforms $\mathbf{T}$ and $\mathbf{U}$ satisfy both the conditions $\mathbf{T}$ is an orthogonal matrix (i.e., $\mathbf{T}^{-1} = \mathbf{T}^T$, where ^{T} denotes transposition) that produces uncorrelated transform coefficients and $\mathbf{U} = \mathbf{T}^{-1}$. Such a transform $\mathbf{T}$ is often called Karhunen-Loève in coding lessons. If the input vector is Gaussian, the optimality of the KLT and its inverse is well known for high rates [11], or when optimal fixed-rate quantizers are employed [13], or more generally when the quantizers are scale-invariant [14] for both the fixed-rate and the variable-

1. They have been freely downloaded from http://eo1.gov/samples.php.

rate coding models. Nevertheless, for non Gaussian sources, it is well known that the KLT can be suboptimal in transform coding [15]. Many papers have been published on optimal transforms in coding associated with various quantizers and entropy coders. Under the high resolution quantization hypothesis, nearly everything is known about the performance of a transform coding system with entropy constrained scalar quantization and mean square distortion. It is then straightforward to find a criterion that, when minimized, gives the optimal linear transform under the above mentioned conditions. Nevertheless, the optimal transform computation is generally considered as a difficult task [16] and the Gaussian assumption is then used in order to simplify the calculus. In [18], the authors resolve the problem of computing the optimal transform (in coding) with mean square error as distortion, under high-rate entropy constraint scalar quantization hypothesis and assuming that $\mathbf{U} = \mathbf{T}^{-1}$. They first prove that such an optimal transform minimizes the criterion

$$\mathcal{C}_1(\mathbf{T}) = I(Y_1; \ldots; Y_N) + \frac{1}{2}\log_2\left(\frac{\det[\mathrm{diag}(\mathbf{T}^{-T}\mathbf{T}^{-1})]}{\det[\mathbf{T}^{-T}\mathbf{T}^{-1}]}\right), \tag{1}$$

where $I(Y_1; \ldots; Y_N)$, $\mathrm{diag}(\mathbf{A})$ and $\det(\mathbf{A})$ denote respectively the mutual information between components of the transformed vector $\mathbf{Y} = (Y_1, \ldots, Y_N)^T$, the diagonal matrix having the same first diagonal as $\mathbf{A}$ (for a square matrix $\mathbf{A}$) and the determinant of $\mathbf{A}$. Then they give two algorithms that minimize the criterion (1), one with no constraint but invertibility of the transform and the other with the orthogonality constraint. Finally, they show on still images that optimal coding transforms returned by their algorithms perform better than any transform usually used in transform coding (DCT, KLT) at high bit-rates (as it was expected) and also at medium and low bit-rates (this was a good surprise). The first term of (1) is a measure of the statistical dependence between the transform coefficients Y_i. It is always non-negative, and zero if and only if the variables are statistically independent. The second term is a pseudo-distance to orthogonality, since it is always non-negative and zero if and only if the column vectors of $\mathbf{T}^{-1}$ are pairwise orthogonal. Furthermore, the criterion is scale invariant: $\mathcal{C}(\mathbf{DT}) = \mathcal{C}(\mathbf{T})$ for any invertible diagonal matrix $\mathbf{D}$.

2. On JPEG2000 optimal spectral transforms

The criterion (1) is not suited for computing an Optimal Spectral Transform (OST) of the JPEG2000 Part2 standard, since the linear transform used to reduce the spectral redundancies is associated with a 2-D DWT applied to each component. Let us introduce the following notations. The encoded hyperspectral (or multi-component) image having N spectral components, each one of dimension $N_r \times N_c$ (N_r rows and N_c columns), is represented by a matrix $\mathbf{X}$ of dimension $N \times L$ (with $L = N_r N_c$) after having scanned each component with a fixed scanning (e.g., row by row from the top to the bottom). The 2-D DWT applied to each component corresponds to the right multiplication by a square matrix $\mathbf{W}^T$ of dimension $L \times L$ and the spectral transform corresponds to the left multiplication by a square matrix $\mathbf{A}$ of dimension $N \times N$:

$$\mathbf{Y} = \mathbf{A}\mathbf{X}\mathbf{W}^T. \tag{2}$$

This compression scheme is called separable, since the spectral transform can be equally applied after or before the 2-D DWT: $\mathbf{A}(\mathbf{X}\mathbf{W}^T) = (\mathbf{A}\mathbf{X})\mathbf{W}^T = \mathbf{Y}$. In addition to the

JEPG2000 standard, it is compatible with the CCSDS standards [19] of spatial image compression. After the entire transformation $\mathbf{X} \mapsto \mathbf{Y}$, a scalar quantizer is applied to each subband of each component. We suppose there are M subbands and $Y_i^{(m)}$ (resp. $\widehat{Y}_i^{(m)}$) denotes any transformed coefficient (resp. dequantized value of $Y_i^{(m)}$) that belongs to the m-th subband of the i-th component. In [9], the authors resolve the problem of computing the optimal spectral transform, when the 2-D DWT is fixed, with mean square error as distortion, under entropy constraint scalar quantization hypothesis and assuming that the decoder applies the inverse transforms to the quantized transform coefficients $\widehat{\mathbf{Y}}$ to get an estimate $\widehat{\mathbf{X}}$ of the original image: $\widehat{\mathbf{X}} = \mathbf{A}^{-1}\widehat{\mathbf{Y}}\mathbf{W}^{-T}$. To be complete, we recall in the next subsection the reasoning that leads to a criterion satisfied by any optimal spectral transform at high-bit rates.

2.1. Expression of the mean-square distortion

It is clear that if $\mathbf{X}$ is a real random vector of dimension N and $\mathcal{A}$ an invertible matrix of order N, if the transformed vector $\mathbf{Y} = \mathcal{A}\mathbf{X}$ is quantized and dequantized in $\widehat{\mathbf{Y}}$ and if the original vector is estimated by $\widehat{\mathbf{X}} = \mathcal{A}^{-1}\widehat{\mathbf{Y}}$, then the end-to-end distortion $D = \frac{1}{N}\,\mathrm{E}[\|\mathbf{X} - \widehat{\mathbf{X}}\|^2]$ satisfies $ND = \mathrm{E}[\mathbf{b}^T\mathcal{A}^{-T}\mathcal{A}^{-1}\mathbf{b}] = \mathrm{tr}[\mathrm{E}(\mathbf{b}\mathbf{b}^T)\mathcal{A}^{-T}\mathcal{A}^{-1}]$, where $\mathbf{b} = \mathbf{Y} - \widehat{\mathbf{Y}}$ is the quantization noise and tr the trace operator. Hence if the hypothesis

$\mathcal{H}_1$: *The components of the quantization noise are zero mean and uncorrelated*

is satisfied (consequently $\mathrm{E}[\mathbf{b}\mathbf{b}^T] = \mathrm{diag}(D_1, \ldots, D_N)$) or if the matrix $\mathcal{A}^{-T}\mathcal{A}^{-1}$ is diagonal (i.e., if the column vectors of $\mathcal{A}^{-1}$ are pairwise orthogonal) then the end-to-end distortion is worth $D = \frac{1}{N}\sum_{i=1}^{N} w_i D_i$ with w_i the i-th column of $\mathcal{A}^{-1}$:

$$w_i = \|\mathcal{A}^{-1}\mathbf{e}_i\|^2 \tag{3}$$

where $\mathbf{e}_i$ is the i-th vector of the canonical basis of $\mathbb{R}^N$.

For the separable compression scheme, the canonical basis of the $N \times L$-matrices' space is the family $(\mathbf{e}_i\mathbf{e}_k'^T)_{i,k}$ ($1 \le i \le N$ and $1 \le k \le L$) where $\mathbf{e}_i$ (resp. $\mathbf{e}'_k$) is the i-th (resp. k-th) vector of the canonical basis of $\mathbb{R}^N$ (resp. $\mathbb{R}^L$) and the relation (3) becomes $\|\mathbf{A}^{-1}\mathbf{e}_i\mathbf{e}_k'^T\mathbf{W}^{-T}\|^2 = \|\mathbf{A}^{-1}\mathbf{e}_i\|^2.\|\mathbf{W}^{-1}\mathbf{e}'_k\|^2$, since the left side of the last equality is the square Euclidean norm of a matrix of rank one. Finally, if we assume that

$\mathcal{H}_3$: *The quantities* $\|\mathbf{W}^{-1}\mathbf{e}'_k\|^2$ *do not depend on the spatial position in the subband*

(this condition is satisfied by DWT using Finite Impulse Response filters when the edge effects are neglected), and introducing $w_i = \|\mathbf{A}^{-1}\mathbf{e}_i\|^2$, $w^{(m)} = \|\mathbf{W}^{-1}\mathbf{e}'_k\|^2$ when $\mathbf{e}'_k$ belongs to the m-th subband ($1 \le m \le M$) and π_m the ratio of column vectors $\mathbf{e}'_k$ that belong to the m-th subband, we have proven that

$$D = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M} \pi_m w_i w^{(m)} D_i^{(m)}, \tag{4}$$

where $D_i^{(m)}$ is the mean-square distortion of the quantizer applied to the m-th subband of the i-th component. It is well known [11] that for a fixed distortion $D_i^{(m)}$ of the quantizer and under the high-resolution quantization hypothesis, the minimum average bit-rate $R_i^{(m)}$ of the quantized value $\widehat{\mathbf{Y}}_i^{(m)}$ (of a transformed coefficient $Y_i^{(m)}$) is achieved

by a scalar quantizer and $R_i^{(m)} \simeq h(Y_i^{(m)}) - \frac{1}{2}\log_2(12D_i^{(m)})$, where $h(Y_i^{(m)})$ is the differential entropy of $Y_i^{(m)}$. Hence the total bit-rate in bits per pixel and per band satisfies $R = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M}\pi_m R_i^{(m)} \simeq \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M}\pi_m[h(Y_i^{(m)}) - \frac{1}{2}\log_2(12D_i^{(m)})]$.

2.2. Criterion minimized by an OST

The high-resolution quantization hypothesis implies condition $\mathcal{H}_1$ and consequently the relation (4). Moreover, the classical optimal bit allocation problem[2] [11] leads to $D_i^{(m)} = \frac{D}{w^{(m)}w_i}$. Injecting this equality in the relation of the total bit-rate, we obtain $R \simeq \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M}\pi_m[h(Y_i^{(m)}) + \frac{1}{2}\log_2(w_i)] + C$, where C does not depend on the spectral transform. We notice that $\prod_{i=1}^{N} w_i = \det \operatorname{diag}(\mathbf{A}^{-T}\mathbf{A}^{-1})$. Let $\mathbf{Z}^{(m)}$ be any column vector of $\mathbf{X}\mathbf{W}^T$ whose components belong to the m-th subband.

We have $\mathbf{A}\mathbf{Z}^{(m)} = \mathbf{Y}^{(m)} = (Y_1^{(m)}, \ldots, Y_N^{(m)})^T$ and since $h(\mathbf{Y}^{(m)}) = \sum_{i=1}^{N} h(Y_i^{(m)}) - I(Y_i^{(m)}; \ldots; Y_N^{(m)})$ and $h(\mathbf{A}\mathbf{Z}^{(m)}) = h(\mathbf{Z}^{(m)}) + \log_2|\det \mathbf{A}|$ ([20]), we have proven that any optimal spectral transform minimizes the criterion

$$\mathcal{C}_2(\mathbf{A}) = \sum_{m=1}^{M}\pi_m I\left(Y_1^{(m)}; \ldots; Y_N^{(m)}\right) + \frac{1}{2}\log_2\left(\frac{\det[\operatorname{diag}(\mathbf{A}^{-T}\mathbf{A}^{-1})]}{\det[\mathbf{A}^{-T}\mathbf{A}^{-1}]}\right). \quad (5)$$

In comparison with the criterion (1), we can remark that the second term is the same (a pseudo-distance to orthogonality of the linear transform) and the first one is now an average per subband of the mutual information between transformed components.

2.3. Optimal Spectral Transforms

In [9], the authors give two algorithms that minimize the criterion (5), one with no constraint but invertibility and the other with the orthogonality constraint. They also show on hyperspectral images that the spectral transforms returned by their algorithms perform better than the KLT at high bit-rates (as expected) and also at medium and low bit-rates and that the performances of the orthogonal OST are very close to the ones of the OST with no constraint. Therefore, in our tests we consider only the orthogonal OST which will be called `OrthOST` in the following. The main drawback of the data-dependent OST is their computational cost which is heavier than the cost of the KLT.

When one disposes of a set of images issued from one spectrometer, it is possible to compute an exogenous OST from a learning basis drawn out of this set. Although it is generally better to use only one OST to reduce the spectral redundancies between all the components of an hyperspectral image rather than several ones, it is not true anymore for exogenous OST. We observed that it is better to split all the components of hyperspectral images in as many blocs as physical sensors and then to learn as much exogenous OSTs as blocs, rather than to learn only one exogenous OST on all the components of the hyperspectral images. This is logical, since one can expect the learning stage adapts the exogenous OST to a physical sensor.

2. Which consists in allocating for each quantizer the value of its distortion $D_i^{(m)}$ in order to minimize the total bit-rate under the constraint of a given end-to-end distortion D.

Table 1. Bit-rate (in bpppb) versus PSNR (in dB) and versus MAD of different spectral transforms on hyperspectral images for the separable scheme.

	PSNR (dB)							MAD						
bit-rate	0.25	0.50	0.75	1.00	1.50	2.00	3.00	0.25	0.50	0.75	1.00	1.50	2.00	3.00
	Learning basis = LB_1 and tested image = *Cuprite*													
Id	54.59	57.39	59.43	61.15	64.16	66.88	72.21	1390	748	510	378	241	163	90
KLT	68.65	70.69	71.90	73.00	75.27	77.74	83.09	512	189	110	87	64	52	28
exo_KLT	68.43	70.36	71.58	72.68	74.95	77.42	82.78	490	197	135	94	69	50	29
OrthOST	69.35	71.20	72.36	73.43	75.70	78.18	83.53	300	160	112	76	61	45	26
exo_OrthOST	68.70	70.87	72.08	73.16	75.41	77.88	83.22	436	160	130	86	64	52	27
	Learning basis = LB_1 and tested image = *Dongting*													
Id	56.53	59.66	62.02	64.01	67.44	70.40	75.77	1300	728	442	285	178	114	61
KLT	68.34	70.87	72.24	73.38	75.65	78.08	83.39	335	149	103	84	66	50	27
exo_KLT	67.85	70.32	71.77	72.95	75.29	77.66	82.95	328	191	112	88	63	53	26
OrthOST	69.18	71.44	72.73	73.86	76.13	78.58	83.88	273	128	92	82	63	47	25
exo_OrthOST	68.33	70.98	72.39	73.55	75.82	78.25	83.53	339	127	99	86	60	48	26
	Learning basis = LB_1 and tested image = *Iranbam*													
Id	53.33	56.62	58.98	60.96	64.39	67.44	72.97	1750	850	616	417	259	165	86
KLT	68.79	71.00	72.31	73.44	75.70	78.16	83.48	435	182	115	87	62	48	27
exo_KLT	68.62	70.72	72.02	73.15	75.43	77.88	83.21	462	224	118	90	65	50	27
OrthOST	69.84	71.78	72.96	74.05	76.32	78.79	84.12	416	156	94	81	63	47	25
exo_OrthOST	68.90	71.31	72.60	73.71	75.95	78.40	83.72	457	195	118	82	68	48	26
	Learning basis = LB_1 and tested image = *Maryland*													
Id	51.20	53.85	55.90	57.72	61.03	64.12	69.94	2392	1176	820	599	392	240	139
KLT	66.26	69.51	71.15	72.40	74.70	77.09	82.35	1250	316	139	104	72	55	32
exo_KLT	65.30	68.80	70.60	71.91	74.23	76.62	81.86	2330	319	179	119	75	55	32
OrthOST	67.05	70.21	71.72	72.91	75.20	77.60	82.87	628	222	104	84	65	49	29
exo_OrthOST	66.31	69.72	71.38	72.62	74.90	77.29	82.55	1000	224	153	88	69	54	29
	Learning basis = LB_1 and tested image = *Tucson*													
Id	53.00	55.56	57.46	59.08	61.93	64.60	69.96	1560	1532	606	478	362	220	122
KLT	66.99	69.57	70.92	72.05	74.30	76.75	82.07	589	206	144	110	71	57	31
exo_KLT	66.77	69.24	70.60	71.73	73.98	76.43	81.77	581	219	148	108	78	57	33
OrthOST	67.84	70.17	71.40	72.49	74.73	77.19	82.53	625	208	122	88	70	53	32
exo_OrthOST	67.06	69.75	71.09	72.20	74.43	76.88	82.19	528	183	126	105	75	60	31

3. Exogenous quasi-optimal spectral transform

In our tests we dispose of ten Hyperion images (*Bay*, *Caledonie*, *Cuprite*, *Dongting*, *Iranbam*, *Maine*, *Maryland*, *Oklahoma*, *Paloalto* and *Tucson*) having 242 spectral bands with 256 columns per band. The number of rows varies from an image to another. First, we keep only 45 bands over the 50 calibrated channels in the visible and near infrared (VNIR) spectrum, constructing in that way ten multi-component images having each $N = 45$ bands and $N_c = 256$ columns. We call them as above and in the following the image *Cuprite* (for example) denotes the $N = 45$ bands image extracted from the original $N = 242$ bands image *Cuprite*. Some images used in our tests are shown in Fig. 1. Then we split these ten images in two disconnected sets, one constituted of P images and which becomes the learning basis, the other constituted of the $10 - P$ remaining images which becomes the test images. Finally the P images of the learning basis are connected band per band and column per column to construct a single virtual large image $\mathbf{X}$ having $N = 45$ bands, $N_c = 256$ columns and a large number of rows. This image is used as input of the previous algorithms which return the exogenous OSTs.

4. Experimental results

We present the bit-rate versus end-to-end distortion obtained with the verification model version 9.1 (VM9 [21]) codec developed by the JPEG2000 group, which uses the Taubman's algorithm EBCOT and the PCRD optimizer applied across components for optimal bit allocation. We considered four distortions: 1) the mean square error (MSE) expressed in terms of the PSNR (Peak Signal to Noise Ratio), $\mathrm{PSNR} = 10 \log_{10} \frac{(2^b-1)^2}{D}$ where D is the actual end-to-end MSE and $b = 16$ is the number of bits per pixel

and per band (bpppb) of initial images; 2) the maximal absolute difference ($\mathrm{MAD} = \max\{|X_i(n) - \widehat{X}_i(n)| : 1 \leq i \leq N \text{ and } 1 \leq n \leq L\}$); 3) the maximum spectral angle

$$\mathrm{MSA} = \max_{1 \leq n \leq L} \left\{ \mathrm{acos} \left(\frac{\sum_{i=1} X_i(n) \widehat{X}_i(n)}{\sqrt{\sum_{i=1}^{N} X_i^2(n) \sum_{i=1}^{N} \widehat{X}_i^2(n)}} \right) \right\} \tag{6}$$

and 4) the mean absolute error ($\frac{1}{NL} \sum_{i=1}^{N} \sum_{n=1}^{L} |X_i(n) - \widehat{X}_i(n)| = \mathrm{MAE}$). With these four distortions, one can estimate the performances of a codec on usual applications of hyperspectral images like classifications and target detection [10].

The results are shown for three different learning bases, each one having $P = 5$ images. We respectively call LB_1, LB_2 and LB_3 the learning bases {*Bay*, *Caledonie*, *Maine*, *Oklahoma*, *Paloalto*}, {*Cuprite*, *Dongting*, *Iranbam*, *Maryland*, *Tucson*} and {*Bay*, *Cuprite*, *Maine*, *Paloalto*, *Tucson*}. Tab. 1 shows the PSNR and the MAD of the five test images associated with LB_1. Tab. 2 shows the average on the five test images of the gain with respect to the KLT of each spectral transform among `OrthOST`, exogenous `OrthOST` (called exo_`OrthOST`) and exogenous KLT (called exo_KLT) for the four distortions. We observe that for the global distortions MSE (PSNR) and MAE, the exogenous `OrthOST` performs always better than the KLT at medium and high bit-rates with an average coding gain that can reach $0.3\,\mathrm{dB}$. And for the local distortions MSA and MAD, the exogenous `OrthOST` performs generally better than the KLT at low, medium and high bit-rates.

5. Conclusion

In many applications using JPEG2000 Part 2 standard, the KLT is generally considered as the best spectral transform for reducing the spectral redundancies of hyperspectral images. However, it is not optimal. In previous works, optimal spectral transforms compatible with the JPEG2000 Part 2 standard have been introduced. They perform better than the KLT, but with a significant heavier computational cost. In this paper, following

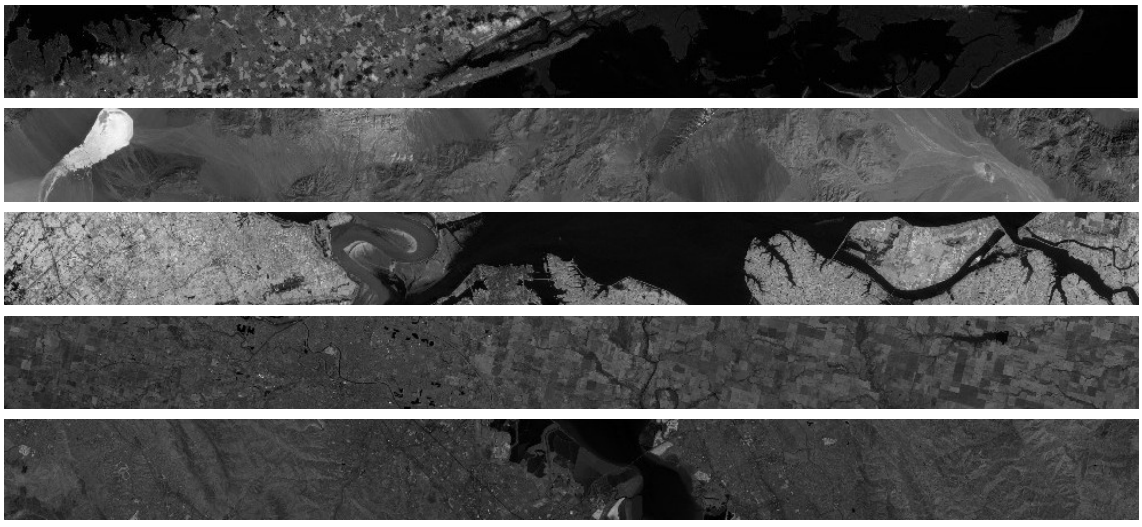


Figure 1. From the top to the bottom: *Bay*, *Cuprite*, *Dongting*, *Oklahoma*, *Paloalto*.

Table 2. Bit-rate (in bpppb) versus 1) average generalized coding gain with respect to the KLT ($\mathrm{PSNR}(\mathbf{A}) - \mathrm{PSNR}(\mathsf{KLT})$), 2) average differences of MAD ($\mathrm{MAD}(\mathbf{A}) - \mathrm{MAD}(\mathsf{KLT})$), 3) average differences of MSA ($\mathrm{MSA}(\mathbf{A}) - \mathrm{MSA}(\mathsf{KLT})$), 4) average differences of MAE ($\mathrm{MAE}(\mathbf{A}) - \mathrm{MAE}(\mathsf{KLT})$).

bit-rate	0.25	0.50	0.75	1.00	1.50	2.00	3.00	0.25	0.50	0.75	1.00	1.50	2.00	3.00
$\mathbf{A}$	average $\mathrm{PSNR}(\mathbf{A}) - \mathrm{PSNR}(\mathrm{KLT})$ (in dB)							average $\mathrm{MAD}(\mathbf{A}) - \mathrm{MAD}(\mathrm{KLT})$						
	Learning basis = LB_1													
exo_KLT	-0.41	-0.44	-0.39	-0.37	-0.35	-0.36	-0.36	214.0	21.6	16.2	5.4	3.0	0.6	0.4
exo_OrthOST	0.05	0.20	0.20	0.19	0.18	0.18	0.17	-72.2	-30.6	3.0	-5.0	0.2	0	-1.2
OrthOST	0.85	0.63	0.53	0.49	0.49	0.50	0.51	-175.8	-33.6	-17.4	-12.2	-2.6	-4.2	-1.6
	Learning basis = LB_2													
exo_KLT	-0.37	-0.32	-0.29	-0.26	-0.24	-0.25	-0.25	-1.2	8.0	18.8	-2.6	5.0	3.4	1.2
exo_OrthOST	-0.10	0.06	0.08	0.09	0.09	0.09	0.07	-124.6	-36.2	-3.6	-11.6	0.8	0.6	0
OrthOST	0.80	0.78	0.65	0.61	0.59	0.59	0.59	-147.4	-67.0	-19.2	-19.4	-5.6	-3.6	-1.2
	Learning basis = LB_3													
exo_KLT	-0.37	-0.40	-0.37	-0.35	-0.35	-0.35	-0.35	74.8	8.4	4.2	5.4	1.8	3.0	0.6
exo_OrthOST	0.30	0.33	0.28	0.26	0.25	0.25	0.25	-150.0	-35.0	-13.2	-7.2	-2.0	-1.6	-1.0
OrthOST	0.88	0.71	0.60	0.56	0.55	0.56	0.57	-188.4	-39.2	-20.4	-12.0	-4.0	-3.6	-1.4
$\mathbf{A}$	average $\mathrm{MSA}(\mathbf{A}) - \mathrm{MSA}(\mathrm{KLT})$ (in $^\circ$)							average $\mathrm{MAE}(\mathbf{A}) - \mathrm{MAE}(\mathrm{KLT})$						
	Learning basis = LB_1													
exo_KLT	0.11	0.10	0.06	0.05	0.04	0.03	0.02	0.98	0.76	0.61	0.51	0.39	0.30	0.16
exo_OrthOST	-0.03	-0.00	0.01	0.01	0.02	0.02	0.01	-0.11	-0.35	-0.30	-0.26	-0.19	-0.14	-0.07
OrthOST	-0.20	-0.03	-0.01	-0.01	0.00	0.01	0.01	-1.80	-1.05	-0.78	-0.65	-0.50	-0.39	-0.22
	Learning basis = LB_2													
exo_KLT	0.13	0.11	0.07	0.04	0.02	0.01	0.01	1.59	0.91	0.66	0.52	0.38	0.30	0.17
exo_OrthOST	-0.05	-0.07	-0.06	-0.02	-0.01	-0.00	-0.00	0.25	-0.23	-0.23	-0.21	-0.16	-0.11	-0.06
OrthOST	-0.14	-0.17	-0.11	-0.07	-0.03	-0.01	-0.01	-1.80	-1.22	-0.87	-0.71	-0.53	-0.41	-0.22
	Learning basis = LB_3													
exo_KLT	0.13	0.11	0.05	0.03	0.01	0.01	0.01	0.90	0.67	0.53	0.44	0.34	0.27	0.14
exo_OrthOST	-0.10	-0.06	-0.05	-0.04	-0.01	-0.01	-0.00	-0.53	-0.50	-0.38	-0.31	-0.23	-0.17	-0.09
OrthOST	-0.20	-0.10	-0.08	-0.06	-0.03	-0.01	-0.01	-1.74	-1.08	-0.79	-0.66	-0.50	-0.39	-0.21

the philosophy given in [7] in replacing the KLT by the orthogonal OST, we computed an exogenous orthogonal OST on a learning basis constituted of hyperspectral images issued from one (and only one) sensor (the spectrometer Hyperion). Then we applied this exogenous quasi optimal transform to other images issued from the same sensor. Using the VM version 9.1 codec of the JPEG2000 group, we compared the performances of this transform with the KLT, the exogenous KLT (obtained with the same philosophy) and the OST, with four different distortions. We have observed that the exogenous OST performs generally better than the KLT at low, medium and high bit-rates.

References

[1] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*, Kluwer Academic, 2002.

[2] D. S. Taubman, "High performance scalable compression with EBCOT", *IEEE Transactions on Image Processing*, Vol. 9, No. 7, pp. 1158–1170, Jul. 2000.

[3] J. T. Rucker, J. E. Fowler and N. H. Younan, "JPEG2000 coding strategies for hyperspectral data", *Proceedings of the International Geoscience and Remote Sensing Symposium*, Vol. 1, pp. 128–131, Jul. 2005.

[4] Q. Du and J. E. Fowler, "Hyperspectral image compression using JPEG2000 and principal component analysis", *IEEE Geoscience and Remote Sensing Letters*, vol. 4, pp. 201–205, Apr. 2007.

[5] B. Penna, T. Tillo, E. Magli and G. Olmo, "Transform coding techniques for lossy hyperspectral data compression", *IEEE Trans. Geoscience and Remote Sensing*, vol. 45, no. 5, May 2007.

[6] E. Christophe, C. Mailhes and P. Duhamel, "Best anisotropic 3D Wavelet decomposition in a rate-distortion sense", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP'06*, vol. II, pp. 17–20, Toulouse (France), May 2006.

[7] C. Thiebaut, E. Christophe, D. Lebedeff, C. Latry, "CNES studies of on-board compression for multispectral and hyperspectral images", *Satellite Data Compression, Communications, and Archiving III Proceedings of the SPIE*, vol. 6683, Edited by R. Heymann, B Huang and I. Gladkova, Sept. 2007.

[8] Q. Du and J. E. Fowler, "Low-Complexity Principal Component Analysis for Hyperspectral Image Compression", *Int. J. High Perform. Comput. Appl.*, vol. 22, no. 4, pp. 438–448, Nov. 2008.

[9] I. P. Akam Bita, M. Barret and D. T. Pham, "Transformations optimales à haut débit pour la compression d'images multi-composnates selon la norme JPEG2000", *Proceedings Colloque GRETSI sur le traitement du signal et des images*, Troyes (France), Sep. 2007.

[10] E. Christophe, D. Léger ans C. Mailhes, "Quality criteria benchmark for hyperspectral imagery", *IEEE Trans. Geoscience and Remote Sensing*, vol. 43, no. 9, pp. 2103–2114, Sep. 2005.

[11] A. Gersho and R. Gray, *Vector Quantization and Signal Compression.* Kluwer, 1992.

[12] V. K. Goyal, "Theoretical foundations of transform coding," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 9–21, 2001.

[13] J.-Y. Huang and P. M. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Commun.*, vol. COM-11, pp. 289–296, Sept. 1963.

[14] V. K. Goyal, J. Zhuang, and M. Vetterli, "Transform coding with backward adaptive updates," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1623–1633, July 2000.

[15] M. Effros, H. Feng, and K. Zeger, "Suboptimality of the Karhunen-Loève transform for transform coding," *IEEE Trans. on Inform. Theory*, vol. 50, no. 8, pp. 1605–1619, 2004.

[16] S. Mallat and F. Falzon, "Analysis of Low Bit Rate Image Transform Coding," *IEEE Trans. Signal Processing*, vol. 46, no. 4, pp. 1027–1042, 1998.

[17] S. Jana and P. Moulin, "Optimality of KLT for high-rate transform coding of Gaussian vector-scale mixtures: application to reconstruction, estimation and classification," *IEEE Trans. Info. Th.* vol. 52, no. 9, pp. 4049–4067, 2006.

[18] M. Narozny, M. Barret and D. T. Pham, "ICA based algorithms for computing optimal 1-D linear block transforms in variable high-rate source coding", *Signal Processing*, vol. 88, no. 2, pp. 268–283, Feb. 2008.

[19] P. S. Yeh, P. Armbruster, A. Kiely, B. Masschelein, G. Moury and C. Schafer, "The new CCSCS image compression recommendation", *Proceedings IEEE Aerospace Conference*, pp. 1–8, Big Sky, Mar. 2005.

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley and Sons, 1991.

[21] *JPEG2000 Verification Model 9.1 (Technical description)*, ISO/IEC JTC 1/SC 29/WG 1 WG1 N2165, Jun. 2001.

Bibliographie

Livre et chapitres de livres en relation avec ma recherche

[1] M. Benidir et M. Barret, *Stabilité des filtres et des systèmes linéaires*, 300 pages, Dunod, 1999.

[2] M. Barret, *chap. 10 : Sur la stabilité des filtres*, du livre *Synthèse de filtres numériques en traitement du signal et des images*, sous la direction de Mohamed Najim, 28 pages, Hermes Lavoisier, 2004. Traduit en anglais : *chap. 10 : Filter stability*, du livre *Digital filters design for signal and image processing*, sous la direction de Mohamed Najim, 22 pages, ISTE Ltd, 2006.

[3] M. Barret, *chap. 11 : Le domaine bi-dimensionnel*, du livre *Synthèse de filtres numériques en traitement du signal et des images*, sous la direction de Mohamed Najim, 45 pages, Hermes Lavoisier, 2004. Traduit en anglais : *chap. 11 : The two-dimensional domain*, du livre *Digital filters design for signal and image processing*, sous la direction de Mohamed Najim, 43 pages, ISTE Ltd, 2006.

Articles dans des revues internationales avec comité de lecture

[4] I. P. Akam Bita, M. Barret et D. T. Pham, "On optimal orthogonal transforms at high bit-rates using only second order statistics in multicomponent image coding with JPEG2000", *Signal Processing*, vol. 90, no. 3, pp. 753–758, mars 2010.

[5] I. P. Akam Bita, M. Barret et D. T. Pham, "On optimal transforms in lossy compression of multicomponent images with JPEG2000", *Signal Processing*, vol. 90, no. 3, pp. 759–773, mars 2010.

[6] H. Bekkouche, M. Barret et J. Oksman, "Adapted generalized lifting schemes for lossless image coding", *Signal Processing*, vol. 88, no. 11, pp. 2790–2803, novembre 2008.

[7] M. Narozny, M. Barret et D. T. Pham, "ICA based algorithms for computing optimal 1-D linear block transforms in variable high-rate source coding", *Signal Processing*, vol. 88, no. 2, pp. 268–283, février 2008.

[8] M. Barret, "Tests de stabilité des filtres numériques récursifs bi-dimensionnels", *Traitement du Signal*, numéro spécial vol. 15, no. 6, pp. 595–602, 1998.

[9] M. Barret et M. Benidir, "Behavior of stability tests for two-dimensional digital recursive filters when faced with rounding errors", *IEEE Transactions on Circuits and Systems (II)*, vol. 44, No. 4, pp. 319–323, avril 1997.

[10] M. Barret et M. Benidir, "On the boundary of the set of Schur polynomials and applications to the stability of 1-D and 2-D digital recursive filters", *IEEE Transactions on Automatic Control*, vol. 39, no. 11, pp. 2335–2339, novembre 1994.

[11] M. Barret et M. Benidir, "A new algorithm to test the stability of 2-D digital recursive filters", *Signal Processing*, vol. 37, no. 2, pp. 255–264, mai 1994.

[12] B. Picinbono et M. Barret, "Nouvelle présentation de la méthode du maximum d'entropie", *Traitement du Signal*, vol. 7, no. 2, pp. 153–158,1989.

Articles dans des conférences internationales avec actes et comité de lecture

[13] I. P. Akam Bita, M. Barret, F. Dalla Vedova et J.-L. Gutzwiller, "Lossy compression of MERIS superspectral images with exogenous quasi optimal coding transforms", *Actes de SPIE Satellite Data Compression, Communications, and Processing V*, San Diego (Californie), 02-06 août 2009.

[14] M. Barret, I. P. Akam Bita, J.-L. Gutzwiller et F. Dalla Vedova, "Lossy hyperspectral images coding with exogenous quasi optimal transforms", *Proceedings of the Data Compression Conference DCC'09*, Snowbird (Utah), 16-18 mars 2009.

[15] J.-L. Gutzwiller, M. Hariti, M. Barret, E. Christophe, C. Thiebaut et P. Duhamel, "Extension du codeur SPIHT au codage d'images hyperspectrales", *Actes de la conférence COmpression et REprésentation des Signaux Audiovisuels CORESA'09*, Toulouse, 19-20 mars 2009.

[16] I. P. Akam Bita, M. Barret, F. Dalla Vedova, J. L. Gutzwiller, "Onboard Compression of Hyperspectral Images Using an Exogenous Orthogonal Quasi-Optimal Transform", *On-Board Payload Data Compression Workshop OBPDC 2008*, Noordwijk (Hollande), 26-27 juin 2008.

[17] I. P. Akam Bita, M. Barret et D. T. Pham, "Transformations optimales à haut débit pour la compression d'images multi-composantes selon la norme JPEG2000", XXIème Colloque GRETSI, pp. 489–492, Troyes (France), 11-14 septembre 2007.

[18] M. Narozny et M. Barret, "ICA-based algorithms applied to image compression", *Proceedings of the IEEE Int. Conf. Acoustic Speech Signal Processing]*, vol. I, pp. 1033–1036, Honolulu (USA), 15-20 avril 2007.

[19] J.-L. Collette et M. Barret, "On simulations about the precision of non uniform hybrid filter bank analog/digital converters", *Proceedings of the IEEE Int. Conf. Acoustic Speech Signal Processing*, Toulouse (France), vol. III, pp. 237–240, 15-19 mai 2006.

[20] I. P. Akam Bita, M. Barret et D. T. Pham, "Compression of multicomponent satellite images using independent component analysis", *Proceedings of the 6th Int. Conf. Independent Component Analysis and Blind Source Separation*, pp. 335–342, Charleston (USA), 5-8 mars 2006.

[21] I. P. Akam Bita, D. T. Pham et M. Barret, "A separation deconvolution modelling for multicomponent images compression", *Workshop on Transform Based on Independent Component Analysis for Audio, Video and Hyperspectral Images Data Reduction and Coding*, Paris, 6–7 juillet 2006.

[22] M. Narozny et M. Barret, "Analyse en composantes indépendantes et compression de données", *Workshop on Transform Based on Independent Component Analysis for Audio, Video and Hyperspectral Images Data Reduction and Coding*, Paris, 6-7 juillet 2006.

[23] M. Narozny, M. Barret, D. T. Pham et I.-P. Akam Bita, "Modified ICA algorithms for finding optimal transforms in transform coding", *4th Int. Symposium on Image and Signal Processing and Analysis*, Zagreb (Croatie), pp. 111–116, 15–17 sept. 2005.

[24] J.-L. Collette, M. Barret, P. Duhamel et J. Oksman, "On hybrid filter bank A/D converters with arbitrary oversampling rate", *4th Int. Symposium on Image and Signal Processing and Analysis*, Zagreb (Croatie), pp. 157–160, 15–17 sept. 2005.

[25] J.-L. Collette, M. Barret et J. Oksman, "Synthesis of hybrid filter banks for A/D conversion : a frequency domain approach", *12th European Signal Processing Conference*, Vienna (Autriche), pp. 109–112, 6–10 septembre 2004.

[26] M. Barret et M. Narozny, "Application of ICA to lossless image coding", *Proceedings of the 4th Int. Conf. Independent Component Analysis and Blind Source Separation*, pp. 855-859, Nara (Japon), avril 2003.

[27] M. Barret, H. Bekkouche, J.-L. Collette et J. Oksman, "Adapted lifting schemes for lossless image coding", *Workshop Transmitting, processing and watermarking multimedia contents*, Bordeaux, pp. 49–54, avril 2003.

[28] H. Bekkouche et M. Barret, "Comparison of lossless codecs for satellite and MRI medical images", vol. II, pp. 475–478, *11th European Signal Processing Conference*, Toulouse France, 3-6 septembre 2002.

[29] H. Bekkouche et M. Barret, "Adaptive mutiresolution decomposition : application to lossless image compression", *Proceedings of IEEE Int. Conf. Acoustic Speech Signal Processing*, vol. IV, pp. 3533–3536, Orlando USA, 13-17 mai 2002.

[30] M. Barret et H. Bekkouche, "Adapted nonlinear multiresolution decomposition with applications in progressive lossless image coding", *Proceedings of 1st Int. Symposium on Image and Signal Processing and Analysis*, Pula, pp. 609–613, juin 2001.

[31] M. Barret et M. Benidir, "An algorithm for robust stability of discret systems", *Proceedings of 8th European Signal Processing Conference*, vol. 3, pp. 1543–1546, Trieste, septembre 1996.

[32] M. Barret et M. Benidir, "Comparison between two stability tests for 2-D digital recursive filters", *Proceedings of 7th European Signal Processing Conference*, vol. 1, pp. 331–334, Edinburg, septembre 1994.

[33] M. Barret et M. Benidir, "Algorithme pour tester la stabilité des filtres digitaux récursifs bi-dimensionnels", *14ème colloque GRETSI*, pp. 1–4, Juan les Pins, septembre 1993.

[34] M. Barret et M. Benidir, "On the optimality of the classical stability criteria for 1-D and 2-D digital recursive filters", *Proceedings of the IEEE Int. Conf. Acoustic Speech Signal Processing*, vol. 3, pp. 65–68, Minneapolis, avril 1993.

[35] M. Benidir et M. Barret, "A Bezout resultant based stability test for 2-D digital recursive filters", *Proceedings of 6th European Signal Processing Conference*, vol. 2, pp. 989–992, Brussels, août 1992.

Posters, communications orales dans des conférences sans comité de lecture

[36] M. Hariti, J. L. Gutzwiller et M. Barret, "Compression d'images hyperspectrales : un codeur par arbres de zéros bien adapté aux transformations à base d'ACI ?", Séminaire Compression d'images multicomposantes du CNES, Toulouse (France), 23 octobre 2008.

[37] M. Barret, D. T. Pham, A. Mohammad-Djafari, P. Duhamel, C. Thiebaut, P. Comon, I. P. Akam Bita, N. Bali, E. Christophe, A. Mohammadpour, M. Narozny et M. Rajih, "Application de l'analyse en composantes indépendantes au codage d'images multicomposantes", *Poster présenté aux journées Panorama des Recherches Incitatives en STIC*, Nancy, 22–24 novembre 2006.

[38] I. P. Akam Bita, M. Barret et D. T. Pham, "Compression of multicomponent satellite images using independent component analysis", *Workshop Geometrical Transforms for Image Processing - Application to Satellite Image Restoration and Compression*, CNES Toulouse, 24 novembre 2005.

[39] M. Barret, D. T. Pham, A. Mohammad-Djafari, P. Duhamel, C. Thiebaut, P. Comon, I. P. Akam Bita, N. Bali, E. Christophe, A. Mohammadpour, M. Narozny et M. Rajih, "Application de l'analyse en composantes indépendantes au codage d'images multicomposantes", *Poster présenté aux journées Panorama des Recherches Incitatives en STIC*, Bordeaux, 21–23 novembre 2005.

[40] M. Barret, D. T. Pham, A. Mohammad-Djafari, P. Duhamel, C. Thiebaut, P. Comon, I. P. Akam Bita, N. Bali, E. Christophe, A. Mohammadpour, M. Narozny et M. Rajih, "Application de l'analyse en composantes indépendantes au codage d'images multicomposantes", *Présentation orale aux Journées de l'ACI Masse de données*, Paris, 16–17 septembre 2004.

[41] H. Bekkouche et M. Barret, "Décomposition multirésolution adaptative. Application à la compression sans perte des images", *Actes de la Réunion des Théoriciens des Circuits de Langue Française*, Paris, 2001.

[42] M. Barret, "Tests de stabilité des filtres numériques", *Actes de la Réunion des Théoriciens des Circuits de Langue Française*, pp. 9–10, Metz, 1999.

Rapports de recherche, prépublications, papiers en préparation

[43] M. Barret, E. Christophe, P. Duhamel, J-L. Gutzwiller, M. Hariti et C. Thiébaut, "Compression d'images multicomposantes par des méthodes à base d'Analyse en Composantes Indépendantes", rapport de recherche, 89 pages, 3 novembre 2008.

[44] M. Barret, "Projet ACI^{2}M", rapport final, 7 pages, 8 décembre 2006.

[45] M. Barret et M. Narozny, "Expression du gain de codage d'une transformation pour de faibles distorsions", rapport de recherche (WP13.ver3) du projet ACI^{2}M, 18 pages, 10 juin 2004.

[46] M. Barret et M. Narozny, "Analyse en composantes indépendantes appliquées au codage d'images multicomposantes", rapport de recherche (WP13.ver2) du projet ACI^{2}M, 12 pages, 31 mars 2004.

[47] M. Barret, "Historique depuis Cauchy jusqu'à Fujiwara des solutions au problème de la localisation des zéros d'un polynôme dans le plan complexe", Prépublication de l'Institut de Recherche de Mathématique Avancée de Strasbourg, pp. 1–56, septembre 1996.

[48] M. Barret, *Étude de la stabilité des filtres numériques récursifs bi-dimensionnels*, thèse de l'Université Paris-Sud, 14 décembre 1993.

[49] *I. P. Akam Bita, M. Barret et D. T. Pham, "Optimal filter banks based on blind separation-deconvolution in variable high-rate source coding", papier long en préparation à soumettre en 2010/2011.*

[50] *M. Barret, J.-L. Gutzwiller et M. Hariti, "A low complexity codec for hyperspectral images", papier long soumis à la revue IEEE Trans. Geoscience and Remote Sensing en nov. 2009.*

[51] *I. P. Akam Bita, M. Barret, F. Dalla Vedova et J.-L. Gutzwiller, "Lossy and lossless compression of MERIS hyperspectral images with exogenous quasi optimal spectral transforms", papier de 14 pages soumis pour publication dans la revue Journal of Applied Remote Sensing.*

Invitation à des séminaires

[52] Séminaire du département de mathématiques de l'Université de Rennes 1, organisé par M.-F. Coste-Roy, sur l'équation de la plus petite hypersurface contenant la frontière de l'ensemble des polynômes de Schur et ses applications sur la localisation des zéros d'un polynôme à N variables par rapport à l'"hyper"-cercle unité, 1997.

[53] Séminaire du département de mathématiques de l'Université de Bretagne Ouest, organisé par P. Saux-Picart, pour présenter les mêmes travaux que ci-dessus, 1996.

[54] Séminaire de l'INRIA Lorraine, organisé par le projet PolKA, sur la stabilité des filtres récursifs bi-dimensionnels et l'ensemble des polynômes de Schur, 1996.

Livre et polycopiés Supélec en relation avec mon enseignement

[55] M. Barret, *Traitement statistique du signal*, 212 pages, Ellipses, 2009.

[56] M. Barret, *Représentation et analyse statistique des signaux : 1 - Définitions, propriétés et modèles de signaux aléatoires*, polycopié de Supélec no. 01224/01a, 80 pages, 2008.

[57] M. Barret, *Ondelettes, analyse multirésolution et décomposition en bancs de filtres*, polycopié de Supélec, 38 pages, 2004.

[58] M. Barret, *Éléments de la théorie des probabilités*, polycopié de Supélec no. 10650, 50 pages, 2007.

[59] M. Barret, *Les fonctions de Bessel*, polycopié de Supélec no. 10400, 44 pages, 1996.

[60] M. Barret, *Rappels et compléments de mathématiques : algèbre*, polycopié de Supélec no. 10600, 97 pages, 1996.

[61] M. Barret, *Rappels et compléments de mathématiques : analyse*, polycopié de Supélec no. 10601, 87 pages, 1996.

[62] M. Barret, *Analyse spectrale*, polycopié no. 10401, 70 pages, 1991.