Univariate and multivariate quantiles, probabilistic and statistical approaches : radar applications

Alexis Decurninge

UPMC PARIS UNIVERSITAS

École Doctorale de Science Mathématiques de Paris Centre

Thèse de doctorat

Discipline : Mathématiques

Spécialité : Statistiques

présentée par

Alexis Decurninge

Quantiles univariés et multivariés, approches probabilistes et statistiques ; applications radar

---

dirigée par Pr. Michel Broniatowski

Co-encadrement industriel : M. Frédéric Barbaresco

JURY

M. Gérard Biau	Université Paris VI	examinateur
M. Paul Deheuvels	Université Paris VI	examinateur
M. Arnaud Guyader	Université Paris VI	examinateur
M. Patrice Bertail	Université Paris X	examinateur
M. Davy Paindaveine	Université Libre de Bruxelles	examinateur
M. François Le Chevalier	University of Delft, Thales	invité
M. Jean-Philippe Ovarlez	Université ONERA	rapporteur
M. Michel Broniatowski	Université Paris VI	directeur
M. Frédéric Barbaresco	Thales	co-directeur

Thèse CIFRE Défense Thales-UPMC

Laboratoire de Statistique Théorique
et Appliquée (LSTA)
UPMC
4, place Jussieu
75252 Paris Cedex 05
Boite courrier 158

UPMC
Ecole Doctorale de Sciences
Mathématiques de Paris Centre
4 place Jussieu
75252 Paris Cedex 05
Boite courrier 290

Remerciements

Paris, le 12 janvier 2015.

Il est grand temps de remercier les personnes qui m'ont soutenu pendant toutes les phases de la production de ce travail.

Tout d'abord, j'aimerais citer M. Barbaresco qui partage volontiers sa passion contagieuse des mathématiques et notamment de la géométrie. Même si nous ne sommes pas arrivés au bout de la "théorie du tout", ses nombreuses références et sa constante motivation à explorer de nouvelles pistes ont été très agréables à vivre.
Je remercie également chaleureusement M. Broniatowski pour avoir été disponible pour de nombreux cafés-idées (deux choses qu'il sait parfaitement agiter) et m'avoir soutenu dans mes propositions. Ses histoires brillamment racontées, sa profonde culture mathématique et sa bonne humeur nous ont presque fait oublier que l'on travaillait.

Je suis honoré que Robert Serfling et Jean-Philippe Ovarlez aient accepté de relire en détail ma thèse. Je vous exprime ici toute ma gratitude ainsi qu'à Paul Deheuvels, Arnaud Guyader, François Le Chevalier, Davy Paindaveine, Patrice Bertail et Gérard Biau pour avoir réussi à dégager du temps dans vos agendas respectifs pour faire partie du jury de cette thèse.

Merci à Fero Matus de m'avoir si gentillement accueilli dans ses bureaux à Prague, à Christian Léonard pour m'avoir donné du grain à moudre sur les transports.

Je remercie également la DGA d'avoir soutenu financièrement cette thèse et, au-delà de l'aspect pécunier, je remercie Philippe Pouliguen, Véronique Serfaty et Philippe Guguen d'avoir pris le temps de suivre les avancements de mes travaux avec régularité.

Je suis infiniment reconnaissant à tous mes compagnons de cellule Svetlana, Patricia, Salim, Soumaya, Petan, Nedjmeddine qui savent tout le bien que je pense d'eux mais aussi à tous les doctorants du LSTA, Cécile, Benjamin, Layal, Amadou, Baptiste, Matthieu, Abdullah, Assia, Roxane, Sarah, Tarn, Mokhtar, Boris... Je n'oublie pas qu'à Thales je pouvais compter sur Paquito, Yann, Christine, Boris, Annelise, Cynthia, Alice et tous les autres doctorants/stagiaires/collègues qui ont agrémentés cette période de cafés, restau et cinémas. Un merci particulier à Claude Adnet pour ses remarques pertinentes sur les problèmes de détection de cibles.

Je remercie également Louise, Corinne et Laurence pour leur disponibilité et leur patience ainsi que Lesley qui a eu le courage d'avoir été la première à lire ce manuscrit écrit en anglais encore approximatif.

Enfin, une pensée pour mes amis et ma famille qui me soutiennent au quotidien et surtout pour Jiao qui aura su m'apporter l'énergie nécessaire à l'accomplissement de ce qui suit et Ulysse qui, un jour peut-être, lira ces lignes.
Merci !

Résumé

Résumé

La description et l'estimation des modèles aussi bien univariés que multivariés impliquant des distributions à queue lourde est un enjeu applicatif majeur. Les L-moments sont devenus des outils classiques alternatifs aux moments centraux pour décrire les comportements en dispersion, asymétrie, kurtosis d'une distribution univariée à queue lourde. En effet, contrairement aux moments centraux correspondants, ils sont bien définis dès que l'espérance de la distribution d'intérêt est finie. Les L-moments peuvent être vus comme la projection de la fonction quantile sur une famille orthogonale de polynômes, récupérant la linéarité inhérente aux quantiles. Nous estimerons dans un premier temps les paramètres de modèles semi paramétriques définis par des contraintes sur ces L-moments par des méthodes de minimisation de divergences.
Nous proposons dans un second temps une généralisation des L-moments aux distributions multivariées qui passe par la définition d'un quantile multivarié défini comme un transport entre la distribution uniforme sur $[0;1]^d$ et la distribution d'intérêt. Cela nous permet de proposer des descripteurs pour des distributions multivariées adaptés à l'étude des queues lourdes. Nous détaillons leurs expressions dans le cadre de modèles possédant des paramètres de rotation.
Enfin, nous proposons des M-estimateurs de la matrice de dispersion des distributions complexes elliptiques. Ces dernières forment un modèle multivarié semi-paramétrique contenant notamment des distributions à queue lourde. Des M-estimateurs spécifiques adaptés aux distributions elliptiques avec une hypothèse supplémentaire de stationnarité sont également proposés. Les performances et la robustesse des estimateurs sont étudiées.
Les signaux radar provenant de fouillis tels les fouillis de mer ou les fouillis de sol sont souvent modélisés par des distributions elliptiques. Nous illustrerons les performances de détecteurs construits à partir de l'estimation de la matrice de dispersion par les méthodes proposées pour différents scénarios radar pour lesquels la robustesse de la procédure d'estimation est cruciale.

Mots-clefs

Distributions à queue lourde, robustesse, M-estimateurs, L-statistiques, L-moments, divergences, méthode de Burg, processus autorégressif stationnaire, quantiles multivariés, transport, descente de gradient sur variété Riemannienne, tessilation, lois elliptiques, modèles semi-paramétriques, inférence, applications radar.

Univariate and multivariate quantiles, probabilistic and statistical approaches; radar applications

Abstract

The description and the estimation of univariate and multivariate models whose underlying distribution is heavy-tailed is a strategic challenge. L-moments have become classical tools alternative to central moments for the description of dispersion, skewness and kurtosis of a univariate heavy-tailed distribution. Indeed, contrary to corresponding central moments, they are well defined since the expectation of the distribution of interest is finite. L-moments can be seen as projections of the quantile function on a family of orthogonal polynomials. First, we will estimate parameters of semi-parametric models defined by constraints on L-moments through divergence methods.
We will then propose a generalization of L-moments for multivariate distributions using a multivariate quantile function defined as a transport of the uniform distribution on $[0;1]^d$ and the distribution of interest. As their univariate versions, these multivariate L-moments are adapted for the study of heavy-tailed distributions. We explicitly give their formulations for models with rotational parameters.
Finally, we propose M-estimators of the scatter matrix of complex elliptical distributions. The family of these distributions form a multivariate semi-parametric model especially containing heavy-tailed distributions. Specific M-estimators adapted to complex elliptical distribution with an additional assumption of stationarity are proposed. Performances and robustness of introduced estimators are studied.
Ground and sea clutters are often modelized by complex elliptical distributions in the field of radar processing. We illustrate performances of detectors built from estimators of the scatter matrix through proposed methods for different radar scenarios.

Keywords

Heavy-tailed distributions, robustness, M-estimators, L-statistics, L-moments, divergence, Burg technique, stationary autoregressive process, multivariate quantile, transport, gradient descent on Riemannian manifold, tessellation, elliptical distributions, semi-parametric models, inference, radar applications.

Table des matières

Introduction

0.1 Considérations générales

Notre guide dans cette thèse a été l'étude et l'estimation des distributions univariées et multivariées dites à queue lourde. La queue d'une loi de probabilité est le comportement de cette loi dans les zones éloignées de sa valeur centrale. Pour les lois univariées sur $\mathbb{R}_+$, le classement des types de queues se fait par rapport à la loi exponentielle de fonction de distribution paramétrée par $\lambda > 0$

$$F_{\mathcal{E}}(x) = (1 - e^{-\lambda x})\mathbb{1}_{x>0}$$

En effet, une loi sur $\mathbb{R}_+$ est dite à queue lourde si sa fonction de distribution F vérifie

$$\int_{\mathbb{R}} e^{\lambda x} dF(x) = \infty \tag{1}$$

pour tout $\lambda > 0$.
Les données simulées sous une distribution appartenant à cette classe ont donc tendance à s'étaler et non à se concentrer autour d'un centre de masse. Autrement dit, les valeurs extrêmes (c'est à dire les valeurs éloignées du centre de masse) ont tendance à être nombreuses pour des lois à queue lourde. C'est pour cela que cela rend leur étude plus complexe que l'étude des lois dites à queue légère dont fait partie la Gaussienne standard notamment.
Ces lois apparaissent naturellement dans plusieurs contextes d'applications, notamment en analyse du risque financier [68], en sciences environnementales [65] et en traitement du signal radar. Nous reviendrons sur cette dernière application qui nous intéresse particulièrement.

Au delà du caractère queue lourde des lois considérées, la robustesse des méthodes employées est un objectif industriel important. En effet, la diversité des scénarios réels est bien souvent difficile à modéliser, les modèles eux-mêmes comme l'estimation des paramètres d'intérêt doivent donc en tenir compte. Nous prendrons comme illustration un échantillon $x_1, ..., x_n$ de variables d'intérêt en faisant l'hypothèse classique que ces échantillons suivent la même loi F_0. Nous citerons deux types de robustesse par rapport à ces hypothèses :

- la robustesse par rapport à une mauvaise spécification du modèle F_0, c'est-à-dire que les données réelles ne vérifient pas les hypothèses du modèle d'étude. Ce genre de considération pousse bien souvent le statisticien à considérer des modèles les plus larges possibles, c'est à dire, à spécifier le modèle avec des a priori non superflus mais suffisants pour décrire le phénomène observé.
- la robustesse par rapport aux valeurs aberrantes pour le modèle. Cette robustesse est différente de la précédente dans le sens où l'hypothèse d'homogénéité des échantillons

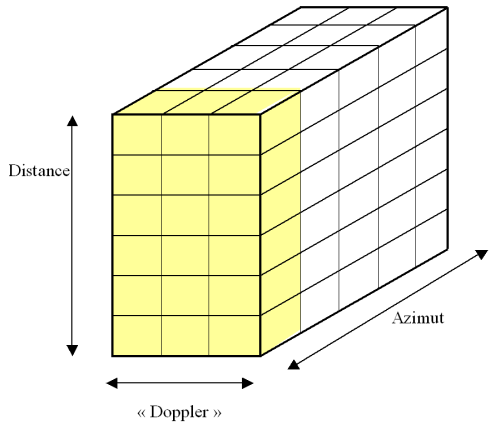


FIGURE 1 – Cube de données radar : l'axe Doppler correspond à l'axe des d impulsions

n'est plus assurée. Les cas extrêmes de contamination par des valeurs aberrantes peuvent aller jusqu'à considérer que $n/2-1$ échantillons ne suivent pas la même loi que les $n/2+1$ autres échantillons.

Nous nous proposons d'étudier des L-statistiques et des M-statistiques robustes dans le cadre des distributions à queue lourde ainsi que l'application de ces dernières pour la détection de cibles radar.

0.2 Contexte d'application : détection de cibles lentes sur fouillis inhomogène

Le contexte applicatif concerne les radars de surfaces (radars de défense aérienne, radars côtiers,...). Les concepts opérationnels d'emploi des radars évoluent pour adapter les nouvelles capacités de détection face à des cibles :

- plus furtives (furtivité passive/active)
- plus lentes ou plus agiles (plus véloces, plus manœuvrantes, à plus basse altitude,...)
- plus intelligentes (antibrouillage, réactivité)
- plus asymétriques (ULM, avions légers...).

Ce type de cibles évoluent dans des environnements fortement perturbés comme les fouillis de sol inhomogènes (cibles à basses altitudes) ou des fouillis de mer très fluctuants et non stationnaires (petites cibles en côtier, missiles sea-skimmer,...).
Les radars que nous considérons envoient des rafales composées de d impulsions dans chaque direction d'espace. Le signal reçu dans chaque direction (azimut) et pour chaque impulsion est découpé en L cases distances (correspondant alors à la portée du radar ; cf Figure 1). Nous nous intéressons plus particulièrement aux angles d'élévation basse (c'est-à-dire aux altitudes basses) où le fouillis est le plus souvent problématique.

Nous traitons chaque azimut de manière indépendante. Pour chaque case d'espace, nous regroupons les d impulsions sous la forme d'un vecteur $x \in \mathbb{C}^d$. Afin d'effectuer un test de détection pour savoir si une cible est présente dans la case d'espace considéré, nous allons prendre en compte les cases environnantes sur l'axe "distance". De telles méthodes adaptatives sont souvent de type CFAR (Constant False Alarm Rate en anglais), c'est à dire que le taux de faux-positifs (la détection à tort d'une cible) doit être maitrisée.

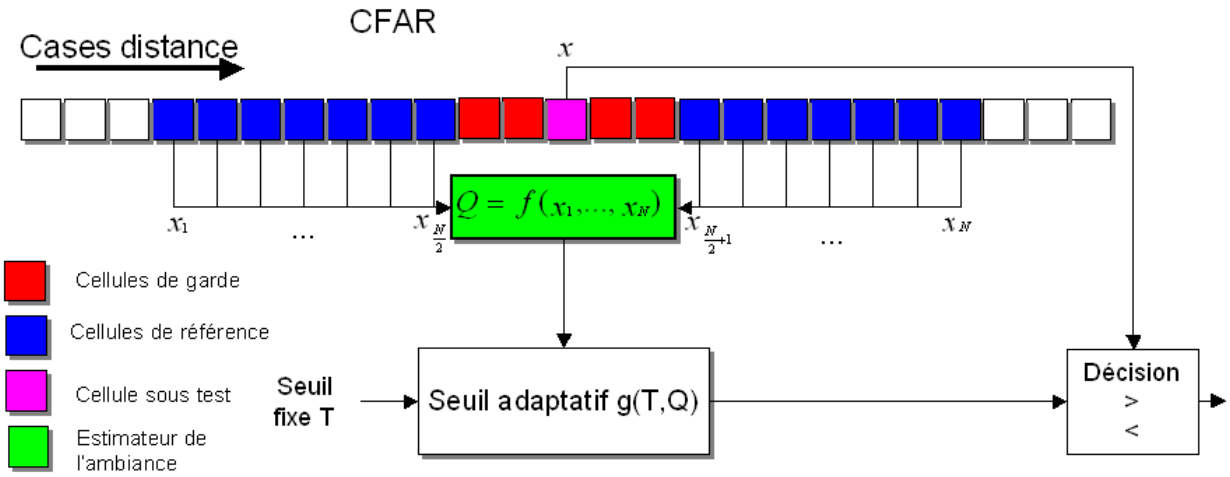


Figure 2 – Algorithme générique CFAR

Tout l'enjeu des algorithmes de détection est donc de modéliser la loi sous-jacente des cases environnantes et de construire un test performant à partir de cette estimation. Nous considérerons :

- la probabilité de détection : la probabilité de détecter une cible si elle est présente dans la case d'intérêt
- la probabilité de fausse alarme : la probabilité de détecter une cible si elle n'est pas présente dans la case d'intérêt

L'objectif principal est de maîtriser la fausse alarme à un taux donné tout en maximisant la probabilité de détection.
La modélisation classique des données environnantes $x_1, ..., x_N \in \mathbb{C}^d$ par une loi Gaussienne complexe n'est pas valable pour les fouillis de sol et de mer qui nous intéressent. Il est bien connu dans la littérature radar que ces lois possèdent une queue de distribution lourde [110][35] (cf Figure 3).

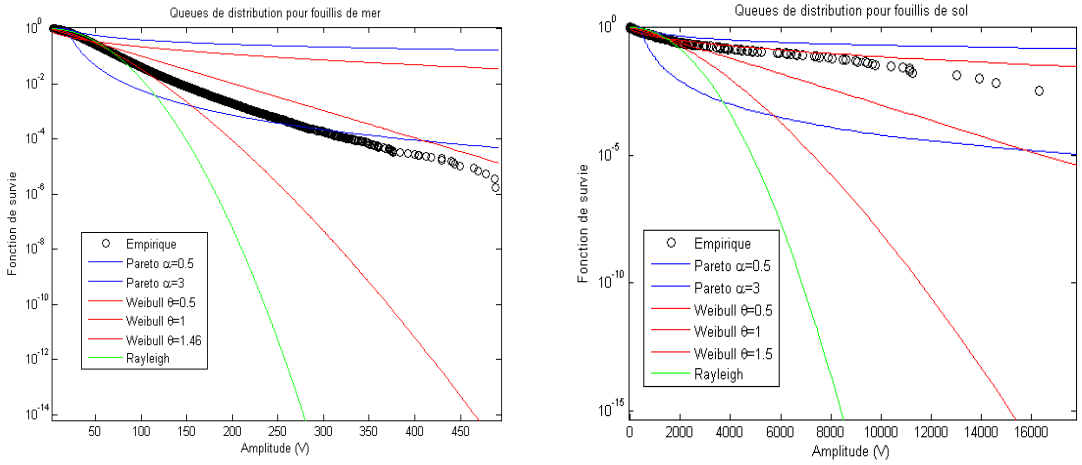
Un modèle semi-paramétrique classiquement considéré pour modéliser les x_i de manière non Gaussienne est le modèle complexe symétrique elliptique. Les distributions elliptiques symétriques complexes centrées sont paramétrées par une matrice de dispersion Σ et une variable aléatoire positive $R \in \mathbb{R}_+$ (nous oublions le paramètre de localisation pour nos applications). En effet, $X \in \mathbb{C}^d$ suit une loi elliptique si et seulement si

$$X \overset{d}{=} R\Sigma^{1/2}U$$

où U suit une loi uniforme sur la sphère unité de $\mathbb{C}^d$. La famille des distributions SIRV (Spherically Invariant Random Variable) est également populaire dans la littérature radar [14][41]. X suit une loi SIRV si et seulement si

$$X \overset{d}{=} R\Sigma^{1/2}Y$$

où Y suit une loi complexe Gaussienne standard sur $\mathbb{C}^d$. Cette seconde famille de lois est en fait comprise dans la famille de lois complexes elliptiques.
La difficulté de l'analyse de ces lois réside dans le fait que la composante non-paramétrique R suit une loi qui dépend de la nature du phénomène observé et possède une distribution



(a) Fouillis de mer (échelle logarithmique)

(b) Fouillis de sol (échelle logarithmique)

FIGURE 3 – Fonctions de survie ($x \mapsto 1 - F(x)$ où F est la fonction de distribution) empirique estimées sur des données réelles d'amplitude de fouillis (i.e. $\|x_i\|$) comparées à des fonctions de survie de familles classiques (distributions Weibull de paramètre de forme θ, distributions Pareto de paramètre de forme α et distribution Rayleigh)

à queue lourde. De plus, nous sommes souvent amené à faire face à une inhomogénéité des fouillis qui implique un besoin en robustesse par rapport aux valeurs aberrantes des algorithmes d'estimation du fouillis. Les deux robustesses évoquées plus haut peuvent être résumées ainsi pour ce modèle elliptique

- une robustesse par rapport à une mauvaise spécification du modèle de R. Il est en effet délicat d'estimer les paramètres d'une loi paramétrique au risque de ne modéliser qu'une partie des cas réels possibles
- une robustesse par rapport à l'inhomogénéité des échantillons. En effet, il est courant que des cibles parasites soient présentes dans les cases environnantes ou qu'un fouillis de type différent (et donc de matrice de dispersion Σ différente) "pollue" les échantillons.

Une hypothèse supplémentaire de stationnarité de second ordre des signaux est souvent faite, notamment pour des rafales courtes (d est faible). Cela implique que la matrice de forme Σ est de type Toeplitz, c'est à dire que ses diagonales sont constantes.

Nous proposerons dans un premier temps des modèles univariés semi-paramétriques définis par des contraintes sur des L-statistiques particulières adaptées à l'étude des distributions à queue lourde : les L-moments. L'idée est de considérer des modèles ne spécifiant que quelques contraintes sans donner de forme particulière à la distribution.
Les L-statistiques sont étroitement liées à la fonction quantile et donc spécifiques aux distributions univariées. Nous proposerons dans une deuxième partie une généralisation des L-moments dans le cas multivarié en les définissant comme des projections d'une fonction quantile multivariée sur une base orthogonale de polynômes.
Enfin, nous proposerons des M-estimateurs de la matrice de dispersion dans le cadre des modèles elliptiques définis ci-dessus. Notre adapterons les méthodes classiques du cas Gaussien pour l'étude des signaux stationnaires au cas elliptique. Dans le cas général non stationnaire, nous définissons deux familles d'estimateurs robustes au sens donné ci-dessus. La performance de ces M-estimateurs sera illustrée dans des cas d'étude liés à la détection

radar.

0.3 Chapitre 1 : estimation pour des modèles définis par des conditions de L-moments

Notons $X_1, ..., X_n$ n variables aléatoires indépendantes et identiquement distribuées (iid) avec une même fonction de distribution F. Alors $X_{1:n} \leq ... \leq X_{n:n}$ les n statistiques d'ordre associées à $X_1, ..., X_n$.
L'existence de ces dernières à partir du moment où l'espérance de la distribution sous-jacente des X_i est finie en fait de bons candidats pour l'étude des distributions à queue lourde. Comme nous le verrons plus loin, l'espérance des statistiques d'ordre peut être vue comme le produit scalaire entre la fonction quantile et un polynôme défniir sur $[0; 1]$. Nous avons donc cherché dans un premier temps à coupler la robustesse inhérente à la fonction quantile avec la robustesse face aux mauvaises spécifications de modèle. En effet, nous cherchons dans le chapitre 1 à étudier les modèles semi-paramétriques constitués par les distributions déterminées par des conditions linéaire sur la fonction quantile (notés modèles SPLQ pour *Semi-Parametric Linear Quantile model*). Ces modèles sont assez larges pour éviter de trop grandes mauvaises spécifications et il contiennent assez d'information pour réaliser une inférence adaptée au phénomène observé.

La description d'une loi univariée passe souvent par l'utilisation des moments centraux : l'espérance, la covariance, l'asymétrie, la kurtosis. Cependant, ces outils font l'hypothèse de l'existence de ces moments, ce qui ne va pas de soi, notamment à cause du caractère queue lourde des distributions que nous considérons. Les L-moments sont une alternative intéressante pour décrire d'une façon similaire les lois univariées en ne faisant l'hypothèse que de l'existence de l'espérance de la loi. Ces outils de description sont donc devenus populaires pour l'étude des lois à queue lourde. Ils sont définis de manière équivalente comme l'espérance de certaines combinaisons linéaires de statistiques d'ordre ou comme produit scalaire de la fonction quantile contre une famille de polynômes orthogonaux dans l'espace des fonction de carré intégrables dans $[0; 1]$.

Soit $r \in \mathbb{N}_* := \mathbb{N} \setminus \{0\}$. Pour des échantillons identiquement distribués $X_1, ..., X_r$, nous notons $X_{1:r} \leq ... \leq X_{r:r}$ ses statistiques d'ordre. Remarquons que $X_{1:r}, ..., X_{r:r}$ sont des variables aléatoires.
Si $\mathbb{E}[|X|] < \infty$, le r-ième L-moment est défini par [64] :

$$\lambda_r(X) = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} \mathbb{E}[X_{r-k:r}] \tag{2}$$

Si nous notons F la fonction de répartition de X et que nous définissons la fonction quantile en $t \in [0; 1]$ comme l'inverse généralisée de F i.e.

$$Q(t) = \inf\{x \in \mathbb{R} \text{ tel que } F(x) > t\},$$

les L-moments peuvent être réécrit :

$$\lambda_r = \int_0^1 Q(t) L_r(t) dt \tag{3}$$

où L_r sont les polynômes de Legendre (translatés sur $[0\,;1]$). La famille de ces polynômes forment une base orthogonale pour le Hilbert $L^2([0; 1], \mathbb{R})$ équipé du produit scalaire usuel

(pour $f, g \in L^2([0;1], \mathbb{R})$, $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$) :

$$L_r(t) = \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k}^2 t^{r-1-k}(1-t)^k = \sum_{k=0}^{r-1} (-1)^{r-k} \binom{r-1}{k}\binom{r-1+k}{k} t^k \tag{4}$$

Les modèles définis par des conditions sur les statistiques d'ordre sont donc composés par l'ensemble des distributions vérifiant ces conditions. Citons deux exemples éclairants.

Example 0.1. *Nous pouvons considérer dans un premier temps l'ensemble des distributions des variables aléatoires X dont les deuxième, troisième et quatrième L-moments vérifient :*

$$\begin{cases} \lambda_2(X) = \sigma(1-2^{-1/\nu})\Gamma(1+1/\nu) \\ \lambda_3(X) = \lambda_2(X)[3 - 2\frac{1-3^{-1/\nu}}{1-2^{-1/\nu}})] \\ \lambda_4(X) = \lambda_2(X)[6 + \frac{5(1-4^{-1/\nu})-10(1-3^{-1/\nu})}{1-2^{-1/\nu}})] \end{cases} \tag{5}$$

avec $\sigma > 0, \nu > 0$. Les L-moments des distributions de ce modèle correspondent aux L-moments d'ordre 2,3,4 d'une loi de Weibull de paramètre σ and ν.

Example 0.2. *Deuxièmement, nous pouvons considérer l'ensemble des distributions des variables aléatoires X vérifiant*

$$\begin{cases} \mathbb{E}[X_{1:3}] = \theta - \nu \\ \mathbb{E}[X_{2:3}] = \theta \\ \mathbb{E}[X_{3:3}] = \theta + \nu \end{cases} \tag{6}$$

pour $\theta \in \mathbb{R}, \nu > 0$. Ce modèle traduit une certaine symétrie des distributions.

Les modèles SPLQ dont les deux exemples précédents font partie sont donc définis par

$$\bigcup_\theta L_\theta = \bigcup_\theta \left\{ F \text{ tel que } \int_0^1 F^{-1}(u)k(u,\theta)du = f(\theta) \right\} \tag{7}$$

avec $\Theta \subset \mathbb{R}^d$, $k : (u, \theta) \in [0;1] \times \Theta \to \mathbb{R}^l$ et $f : \Theta \to \mathbb{R}^l$.

Dans le cadre des contraintes de L-moments, ce modèle se réécrit

$$\bigcup_\theta L_\theta = \bigcup_\theta \left\{ F \text{ s.t. } \int_0^1 L(u)F^{-1}du = f(\theta) \right\} \tag{8}$$

avec L la concaténation des L-moments considérés

$$k(u,\theta) = L(u) := \begin{pmatrix} L_1(u) \\ \vdots \\ L_l(u) \end{pmatrix}.$$

Par analogie avec les modèles définis par des équations de moment, nous définirons des procédures d'estimation basées sur un critère de minimisation de divergences entre la distribution empirique et l'ensemble des distributions du modèle. Cependant, ces modèles ne jouissent pas d'une linéarité par rapport à la fonction de distribution mais par rapport à la fonction quantile. Or, la linéarité des contraintes par rapport au critère à minimiser est un argument essentiel pour la simplicité de la mise en œuvre de ces méthodes d'estimation. Ainsi, nous proposons de considérer les divergences entre mesures quantiles et

non entre mesures de probabilités.

Pour ce faire, nous introduisons donc les φ-divergences indépendamment introduites par Csiszar [42] et Ali and Silvey [2] dans le contexte des mesures de probabilités. Ces divergences sont également valables pour définir une notion de "distance" entre mesures σ-finies.
Soit $\varphi : \mathbb{R} \to [0, +\infty]$ une fonction strictement convexe avec $\varphi(1) = 0$ de domaine $dom(\varphi) := (a_\varphi, b_\varphi)$. Si dF et dG sont deux mesures σ-finies de $(\mathbb{R}, B(\mathbb{R}))$ telles que dG est absolument continue par rapport à dF, nous définissons la φ-divergence entre dF et dG par :

$$D_\varphi(G, F) = \int_{\mathbb{R}} \varphi\left(\frac{dG}{dF}(x)\right) dF(x) \tag{9}$$

où $\frac{dG}{dF}$ désigne la dérivée de Radon-Nikodym entre dF et dG. Une propriété importante est que

$$D_\varphi(G, F) = 0 \text{ si et seulement si } F = G.$$

Notons que cette divergence n'est pas une distance dans le sens où la propriété classique de symétrie des distances n'est pas respectée.
Soit $x_1, ..., x_n$ un échantillon iid issue d'une distribution F. L'estimateur que nous proposons est alors naturellement défini comme la minimisation de la φ-divergence entre le modèle et le quantile empirique. Comme l'hypothèse d'absolue continuité de dG par rapport à dF est indispensable pour que les φ-divergences, il est inévitable de réduire le modèle aux distributions dont le quantile est absolument continu par rapport à la mesure associée au quantile empirique.

$$\hat{\theta}_n = \arg\inf_{\theta\in\Theta} \inf_{G^{-1}\ll F_n^{-1}; \int_0^1 L(u)G^{-1}(u)du = -f(\theta)} \int_0^1 \varphi\left(\frac{dG^{-1}}{dF_n^{-1}}(u)\right) dF_n^{-1}(u) \tag{10}$$

$$= \arg\inf_{\theta\in\Theta} \inf_{y:=(y_1,...,y_n); \sum_{i=1}^{n-1} K(i/n)(y_{i+1}-y_i) = f(\theta)} \sum_{i=1}^{n-1} \varphi\left(\frac{y_{i+1}-y_i}{x_{i+1}-x_i}\right)(x_{i+1}-x_i) \tag{11}$$

En effet, nous pouvons caractériser les L-moments d'ordre supérieur à 2 grâce à l'égalité suivante

$$-\mathbb{E}\left[\frac{1}{r}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}X_{k:n}\right] = \int_0^1 K_r(t)dF^{-1}(t) = f_r(\theta). \tag{12}$$

Un peu de travail supplémentaire nous offre une réécriture de l'estimateur $\hat{\theta}_n$ ci-dessus comme une minimisation d'une "énergie de déformation" de la distribution empirique sans condition d'absolue continuité

$$\hat{\theta}_n = \arg\inf_{\theta\in\Theta} \inf_{T\in L''_\theta(F_n)} \int_{\mathbb{R}} \varphi\left(\frac{dT}{d\lambda}\right) d\lambda, \tag{13}$$

avec

$$L''_\theta(F) = \left\{T : \mathbb{R} \to \mathbb{R} \text{ tel que } \int_{\mathbb{R}} K(F(x))dT(x) = f(\theta)\right\}. \tag{14}$$

Le modèle s'écrit alors

$$\left\{F \circ T^{-1} \in M \text{ tel qu'il existe } \theta \text{ tel que } \int_{\mathbb{R}} K(F \circ T^{-1}(x))dx = f(\theta)\right\}$$
$$= \left\{F \circ T^{-1} \in M, T \in \cup_{\theta\in\Theta} L'_\theta(F)\right\}.$$

où M est l'ensemble des mesures de mesure totale 1.

Notons que $\cup_\theta L''_\theta(F)$ représente le modèle défini par des contraintes de L-moment vu à travers une mesure de référence F. Il peut être vu comme l'espace des mesures G satisfaisant les conditions de L-moments telle que G est la déformation de la mesure de référence par T.

Notons λ la mesure de Lebesgue définie sur $\mathbb{R}$. Sous des hypothèses générales, nous obtenons une formule de dualité pour l'estimateur ci-dessus

Theorem 0.1. *Supposons qu'il existe $T \in L''_\theta(F)$ tel que $a_\varphi < \frac{dT}{d\lambda} < b_\varphi$ λ-p.s. alors*

$$\inf_{T \in L''_\theta(F)} \int_{\mathbb{R}} \varphi\left(\frac{dT}{d\lambda}\right) d\lambda = \sup_{\xi \in \mathbb{R}^l} \langle \xi, f(\theta) \rangle - \int_{\mathbb{R}} \psi(\langle \xi, K(F(x)) \rangle) d\lambda \tag{15}$$

Ce résultat nous permet de transformer un problème d'optimisation défini sur un espace de dimension infinie et sous contraintes en un problème d'optimisation non contraint en dimension finie.
Ceci nous permettra également prouver des résultats de consistance et de normalité asymptotique de l'estimateur $\hat{\theta}_n$ définie par l'équation 13 ci-dessus sous des hypothèses classiques.

0.4 Chapitre 2 : L-moments multivariés

Nous nous attacherons dans un deuxième temps à proposer une généralisation des L-moments pour la description des lois multivariées dans $\mathbb{R}^d$.
Les L-moments définis plus haut possèdent des propriétés intéressantes que nous voudrions préserver pour définir leurs versions multivariées. Serfling et Xiao [99] ont listé les caractéristiques des L-moments univariés

- Leur existence à tous les ordres à partir du moment où l'espérance est finie
- Une distribution est caractérisée par ses L-moments
- Une représentation en tant que produit scalaire contre une famille de fonctions orthogonales
- Une représentation en tant qu'espérance d'une L-statistique (i.e. une fonction linéaire par rapport aux échantillons)
- Les L-moments empiriques peuvent être vus comme des U-statistiques ce qui permet l'obtention de résultats asymptotiques
- Les L-moments empiriques sont également des L-statistiques ce qui permet une estimation rapide
- L'existence d'une version empirique des L-moments non biaisée vue soit comme une U-statistique ou une L-statistique
- Les L-moments empiriques sont plus stables que les moments centraux, cette stabilité croissant avec l'ordre des moments : l'impact de chaque grande valeur x est linéaire pour les L-moments alors qu'elle est de l'ordre de $(x-\bar{x})^k$ pour les moments classiques d'ordre k.

Nous ajoutons à cette liste deux propriétés supplémentaires

- L'équivariance des L-moments par rapport aux homothéties et leur invariance par rapport aux translations si l'ordre est supérieur à deux
- La maniabilité des L-moments pour la plupart des familles paramétriques classiques ce qui rend la méthode des L-moments intéressante dans le cadre de l'estimation paramétrique, notamment pour l'estimation des paramètres de queue d'une loi à

queue lourde.

Soit $X = (X_1, ..., X_d)^T \in \mathbb{R}^d$ un vecteur aléatoire. Serfing et Xiao ont proposé une extension multivariée des L-moments de X en définissant ce qu'ils appellent des L-comoments construits à partir des distributions conditionnelles de chaque couple $(X_i, X_j)^T \in \mathbb{R}^2$ pour $i, j \in \{1, ..., d\}$. Les L-comoments conservent la majorité des propriétés des L-moments univariés listées ci-dessus à l'exception notable de la caractérisation d'une distribution multivariée par ses L-moments.
Nous proposons de généraliser leur approche par un léger changement de point de vue. Afin de conserver la caractérisation d'une distribution par ses L-moments, il semble inévitable d'abandonner l'existence d'une représentation en tant que U-statistiques (et donc l'accès simple à une version non biaisée des L-moments).

Notre point de départ pour définir des L-moments multivariés consiste en leur caractérisation 3 en tant que projections orthogonales d'une fonction quantile sur une base orthogonale de polynômes définis sur $[0; 1]$. En effet, il est facile de définir des polynômes orthogonaux sur le cube unité $[0; 1]^d$ qui apparaît alors comme la généralisation naturelle de l'espace $[0; 1]$. Nous allons donc proposer une définition d'un quantile multivarié comme fonction de $[0; 1]^d$ dans $\mathbb{R}^d$.

Comme il n'y a pas d'ordre total dans $\mathbb{R}^d$, il n'y a pas une unique manière de définir des quantiles multivariés. Nous choisissons une approche qui utilise la notion de transport de mesures. Cela est naturel car, en dimension un, la fonction quantile est un transport de la mesure uniforme sur $[0; 1]$ vers la mesure d'intérêt. Par exemple, Galichon et Henry [54] ont proposé de définir les quantiles multivariés comme les transport optimaux quadratiques (voir par exemple [108] pour la définition d'un transport optimal) de la mesure uniforme sur $[0; 1]^d$ vers la mesure multivariée qui nous intéresse. Nous reprenons cette définition en relaxant l'hypothèse d'optimalité du transport et ainsi pouvons définir les L-moments multivariés comme projections de ce transport sur une famille de polynômes définis sur $[0; 1]^d$.

Il est donc temps de définir un transport $T : \mathbb{R}^d \to \mathbb{R}^d$ entre deux mesures μ et ν définies sur $\mathbb{R}^d$.

Definition 0.1. *La mesure image μ par T est la mesure notée $T\#\mu$ satisfaisant*

$$T\#\mu(B) = \mu(T^{-1}(B)) \text{ pour tout Borélien } B \text{ de } \mathbb{R}^d \tag{16}$$

T est un transport de μ vers ν si $T\#\mu = \nu$. On dit que μ est la mesure source et ν la mesure cible. De plus, si X et Y sont deux vecteurs aléatoires de mesure respective μ et ν, alors $T(X) \overset{d}{=} Y$

Ainsi, nous définissons les L-moments multivariés comme suit

Definition 0.2. *Soit $Q : [0; 1]^d \to \mathbb{R}^d$ un transport de la mesure uniforme sur $[0; 1]^d$ vers ν. Alors, si $\mathbb{E}[\|X\|] < \infty$, le L-moment de multi-indice $\alpha \in \mathbb{N}_*^d$ associé au transport Q est :*

$$\lambda_\alpha = \int_{[0;1]^d} Q(t_1, ..., t_d) L_\alpha(t_1, ..., t_d) dt_1 ... dt_d \in \mathbb{R}^d \tag{17}$$

où la famille des L_α est la famille orthogonale complète des polynômes de Legendre multivariés

$$L_\alpha(t_1, ..., t_d) = \prod_{k=1}^{d} L_{i_k}(t_k).$$

Les polynômes L_{i_k} sont les polynômes de Legendre univariés définies par 4.

Il existe plusieurs manières de transporter une mesure sur une autre. Nous considérerons deux familles de transport

- le transport de Rosenblatt
- les transports optimaux

Si nous considérons comme quantile associé à un vecteur aléatoire X le transport de Rosenblatt de la mesure uniforme sur $[0;1]^d$ vers la mesure associée à X (qui est un transport basé sur les distributions conditionnelles successives des composantes de X), une sous-famille des L-moments qui en découlent correspond aux L-comoments de Serfling et Xiao [99].

Nous considérerons une deuxième construction des quantiles basés sur les transports optimaux quadratiques que nous appellerons transport monotones.
Soit $[0;1]^d$ le cube unité de $\mathbb{R}^d$ et soit $\mathcal{N}_d$ la mesure Gaussienne canonique définie sur $\mathbb{R}^d$. La fonction $Q_0 : [0;1]^d \to \mathbb{R}^d$ définie par

$$Q_0(t_1, ..., t_d) = \begin{pmatrix} \mathcal{N}_1^{-1}(t_1) \\ \vdots \\ \mathcal{N}_1^{-1}(t_d) \end{pmatrix} \tag{18}$$

transporte la mesure uniforme sur $[0;1]^d$ vers $\mathcal{N}_d$. Ceci définit une mesure de référence $\mathcal{N}_d$.

Notons $\mu = \mathcal{N}_d$ et ν une mesure d'intérêt sur $\mathbb{R}^d$. Avec T un transport optimal de μ vers ν, nous définissons un quantile de X, vecteur aléatoire de mesure ν, par

$$Q := T \circ Q_0 \tag{19}$$

La mesure de référence Gaussienne n'est pas nécessaire et nous pouvons définir un quantile multivarié comme un transport optimal entre la mesure uniforme sur $[0;1]^d$ et la mesure d'intérêt ν sur $\mathbb{R}^d$. L'intérêt de cette mesure de référence μ réside dans le fait que le transport T de μ vers ν est plus facile à définir si ν appartient à une certaine classe de distributions contenant des paramètres de rotation, telles la famille des distributions elliptiques, sur lesquelles nous allons porter notre attention.
Même si, à notre connaissance, il n'y a pas de formes closes d'un transport monotone de la distribution uniforme sur $[0;1]^d$ (ou même de la Gaussienne standard multivariée) vers une distribution elliptique, nous pouvons définir des alternatives permettant de prendre en compte des paramètres de même type.

Le prix à payer pour utiliser les transports monotones est souvent d'abandonner les modèles classiques et de considérer des modèles construits à partir du transport. En effet, une façon naturelle d'utiliser ces quantiles est de définir les modèles non plus à partir d'une fonction densité mais à partir de leur quantile.
Nous avons brièvement présenté différentes façons de définir un quantile multivarié comme transport de la distribution uniforme sur $[0;1]^d$ vers ν. Ainsi, à chaque définition d'un quantile multivariée correspond une définition des L-moments associés à ce quantile. Dans ce chapitre, nous analyserons les qualités d'invariance/équivariance des L-moments introduits ainsi que les propriétés de consistence des estimateurs plug-in associés à chaque version des L-moments multivariés.

0.5 Chapitre 3 : M-estimateur pour des modèles elliptiques

Dans un premier temps, nous nous pencherons sur l'estimation de la matrice de dispersion Σ d'une distribution elliptique dans le cas stationnaire. Sa forme Toeplitz nous permet de décomposer l'estimation de Σ (de taille $d \times d$) en d estimations de matrices Toeplitz de taille 2×2. Ce découpage correspond à une application de la "technique de Burg" [31]. Dans le cas Gaussien, au lieu d'estimer directement la matrice de covariance d'un vecteur aléatoire $Y \in \mathbb{C}^d$, nous définissons itérativement des vecteurs aléatoires intermédiaires à partir de Y, éléments de $\mathbb{C}^2$, dont la matrice de covariance théorique peut être exprimée en fonction de Σ.
La méthode itérative d'estimation des matrices covariances de taille 2×2 a été originellement proposée pour l'estimation des paramètres d'une série temporelle Gaussienne autorégressive stationnaire. Or, Y peut être vu comme la trace d'une série temporelle de taille d. L'analogie entre le processus autorégressif et sa trace permet donc d'adapter les outils des processus autorégressifs pour l'estimation de Σ. De plus, considérer Y comme la trace d'une processus Gaussien autorégressif d'ordre $M < d-1$ permet de considérer une structure supplémentaire pour la matrice Σ (plus contraignante que la structure Toeplitz).
Nous proposons dans un premier temps d'adapter les méthodes de Burg d'estimation de la covariance d'un vecteur Gaussien à des mélanges multiplicatifs de vecteurs autorégressifs définis par

$$X = RY \in \mathbb{C}^d \tag{20}$$

où $R > 0$ représente la composante scalaire à queue lourde. La famille des vecteurs Y est donc une sous-famille de modèles SIRV.
Si nous notons $a_1^{(M)}, ..., a_M^{(M)}$ les paramètres du modèle autorégressif sous-jacent à Y, nous avons

$$Y_n + \sum_{i=1}^{M} a_i^{(M)} Y_{n-i} = b_n \tag{21}$$

pour $n \leq d$. Alors X suit lui aussi le même type d'autorégression avec des innovations non-Gaussiennes

$$X_n + \sum_{i=1}^{M} a_i^{(M)} X_{n-i} = Rb_n \tag{22}$$

Nous adaptons alors, pour les échantillons du vecteur X, les méthodes de Burg classiques afin de les rendre indépendantes de la loi de la variable aléatoire R.
L'inconvénient de telles méthodes est leur non-robustesse face aux valeurs aberrantes. Nous proposons alors des variantes de ces estimateurs par des méthodes de type médian. En effet, notre idée est de calculer le médian d'estimation de la matrice de dispersion sur des sous-échantillons de $y_1, ..., y_N$ de taille S donnée.

Dans le cas général de l'estimation de la matrice de dispersion pour des échantillons tirés selon une loi elliptique, Maronna [78] a proposé des M-estimateurs robustes de Σ de type Huber satisfaisant l'équation :

$$\Sigma = \frac{1}{N} \sum_{i=1}^{N} u(x_i^+ \Sigma^{-1} x_i) x_i x_i^+. \tag{23}$$

La fonction u doit satisfaire certaines conditions pour que l'estimateur ainsi défini soit bien défini et consistant. Le défaut principal de ces estimateurs est leur non-invariance

par rapport à la distribution de l'amplitude R. Ainsi, Tyler [104] a proposé l'estimateur suivant dans le cas réel

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}\frac{x_i x_i^+}{x_i^+\Sigma^{-1}x_i}. \tag{24}$$

La fonction $u(x) = \frac{1}{x}$ ne satisfait pas les conditions de Maronna mais Tyler a montré que cet estimateur satisfait des propriétés de consistance. Il a montré également que c'est le maximum de vraisemblance des échantillons normalisés $\frac{x_1}{\|x_1\|}, .., \frac{x_N}{\|x_N\|}$ (appelés signes ou angles multivariés). Comme la plupart des estimateurs de maximum de vraisemblance, l'estimateur de Tyler ne présente que peu de robustesse face aux valeurs aberrantes.
Nous proposons d'adapter l'approche de Huber par M-estimateurs (cf Huber and Ronchetti [66]) aux échantillons normalisés $\frac{x_1}{\|x_1\|}, .., \frac{x_N}{\|x_N\|}$. Le vecteur aléatoire $\frac{X}{\|X\|}$ suit une distribution ACG (pour Angular Central Gaussian en anglais) [105]. Sa densité est donnée par

$$f_\Sigma(x) = \frac{\Gamma(d)}{2\pi^d|\Sigma|}(x^+\Sigma^{-1}x)^{-d}\mathbb{1}_{\|x\|=1} \tag{25}$$

Nous combinons ainsi la robustesse aux valeurs aberrantes avec une invariance de notre processus d'estimation de Σ par rapport à la distribution de l'amplitude R. Les M-estimateurs que nous considérons sont liées à deux types de divergences différents des φ-divergences définies au premier chapitre.
Premièrement, nous considérons des M-estimateurs tirés d'une ψ-divergence définies entre deux densités f et g :

$$d_\psi(f,g) = \int \psi(\log g(z))g(z)dz - \int \psi(\log f(z))g(z)dz + \int \Psi^*(\log f(z))dz - \int \Psi^*(\log g(z))dz. \tag{26}$$

Cela implique un estimateur donné par

$$\hat{\Sigma} = \arg\min_\Sigma -\frac{1}{N}\sum_{i=1}^{N}\psi(\log f_\Sigma(z_i)) + \int \Psi^*(\log f_\Sigma(z))dz. \tag{27}$$

Nous comparerons ces dernières avec des γ-divergences

$$d_\gamma(f,g) = \log\left[\frac{\left(\int g(x)^{1+\gamma}dx\right)^{1/\gamma(1+\gamma)}\left(\int f(x)^{1+\gamma}dx\right)^{1/(1+\gamma)}}{\left(\int g(x)f(x)^\gamma dx\right)^{1/\gamma}}\right] \tag{28}$$

qui nous donnent l'estimateur suivant

$$\hat{\Sigma} = \arg\min_\Sigma -\frac{1}{\gamma}\log\left[\frac{1}{N}\sum_{i=1}^{N}f_\Sigma(x_i)^\gamma\right] + \frac{1}{\gamma+1}\log\left[\int f_\Sigma(x)^{1+\gamma}dx\right] \tag{29}$$

Nous donnerons des résultats de consistance de ces M-estimateurs et illustrerons leur robustesse ainsi que celles des estimateurs de type Burg définis dans le cas stationnaire face à plusieurs scénarios radar.

Chapitre 1

Estimation under L-moment condition models

1.1 Motivation and notation

For univariate distributions, L-moments are expressed as the expectation of a particular linear combination of order statistics. Let us consider r independent copies $X_1, ..., X_r$ of a random variable X with $\mathbb{E}(|X|)$ a finite number. The r-th L-moment is defined by

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} \mathbb{E}[X_{r-k:r}] \tag{1.1}$$

where $X_{1:r} \leq ... \leq X_{r:r}$ denotes the order statistics. The four first L-moment can be considered as a measure of location, dispersion, skewness and kurtosis. Indeed $\lambda_1 = \mathbb{E}(X)$, λ_2 is expressed as $\lambda_2 = (1/2)\,\mathbb{E}(|X - Y|)$ with Y an independent copy of X, λ_3 indicates the expected distance between the mean of the extreme terms and the median one in a sample of three i.i.d. replications of X, and λ_4 is an indicator of the expected distance between the extreme terms of a sample of four replicates of X with respect to a multiple of the distance between the two central terms.

L-moments constitute a robust alternative to traditional moments as descriptors of a distribution since only the existence of $\mathbb{E}(|X|)$ is needed in order to insure their existence. Since their introduction in Hosking's paper in 1990 ([64]), methods based on L-moments have become popular especially in applications dealing with heavy-tailed distributions. As mentioned in [64] and [**?**] :"The main advantage of L-moments over conventional moments is that L-moments, being linear functions of the data, suffer less from the effect of sampling variability : L-moments are more robust than conventional moments to outliers in the data and enable more secure inferences to be made from small samples about an underlying probability distribution. Also as seen through (1.1) the L-moments are determined by the expectation of extreme order statistics, and vice versa". This motivates their success for the inference in models pertaining to the tail behavior of random phenomenons.

In this chapter, we will consider semi-parametric models conditioned by constraints on a finite number of L-moments. Let us mention three examples of such models ; the two first examples describe neighborhoods of the Weibull and the Pareto models, which are classical benchmarks for the description of tail properties, and the third one describes a family of distributions which express some loose symmetry property.

Example 1.3. *We first consider the model which is the family of all the distributions of a r.v. X whose second, third and fourth L-moments verify :*

$$\begin{cases} \lambda_2 = \sigma(1-2^{-1/\nu})\Gamma(1+1/\nu) \\ \lambda_3 = \lambda_2[3-2\frac{1-3^{-1/\nu}}{1-2^{-1/\nu}})] \\ \lambda_4 = \lambda_2[6+\frac{5(1-4^{-1/\nu})-10(1-3^{-1/\nu})}{1-2^{-1/\nu}})] \end{cases} \tag{1.2}$$

for any $\sigma > 0, \nu > 0$. These distributions share their first L-moments of order 2, 3 and 4 with those of a Weibull distribution with scale and shape parameter σ and ν. When X is substituted by $Y := X + a$ for some real number a then the distribution of Y is Weibull with a shifted support, hence with the same parameters σ and ν as X ; the r.v. Y shares the same L-moments λ_r with those of X but for $r = 1$ and the model (1.2) describes a neighborhood of the continuum of all Weibull distributions on $[a, \infty)$ or on $(-\infty, a]$ when a belongs to $\mathbb{R}$. Hence this model aims at describing a shape constraint on the tail of the distribution of the data, independently of its location.

Example 1.4. *Secondly, we consider the model which is the space of the distributions whose second, third and fourth L-moments verify :*

$$\begin{cases} \lambda_2 &= \frac{\sigma}{(1-\nu)(2-\nu)} \\ \lambda_3 &= \lambda_2 \frac{1+\nu}{3-\nu} \\ \lambda_4 &= \lambda_2 \frac{(1+\nu)(2+\nu)}{(3-\nu)(4-\nu)} \end{cases} \tag{1.3}$$

for any $\sigma > 0, \nu \in \mathbb{R}$. These distributions share their first L-moments with those of a generalized Pareto distribution with scale and shape parameter σ and ν. The same remark as in the above example holds ; model (1.3) describes a neighborhood of the whole continuum of Pareto distributions on $[a, \infty)$ or on $(-\infty, a]$ when a belongs to $\mathbb{R}$.

Example 1.5. *Let finally be given an appealing example based on order statistics, namely*

$$\begin{cases} \mathbb{E}[X_{1:3}] = \theta - \nu \\ \mathbb{E}[X_{2:3}] = \theta \\ \mathbb{E}[X_{3:3}] = \theta + \nu \end{cases}$$

for any $\theta \in \mathbb{R}, \nu > 0$.

Before any further discussion on the scope of the present paper, a few notation seems useful. For a non decreasing function F with bounded variation on any interval of $\mathbb{R}$ we denote $\mathbf{F}$ the corresponding positive $\sigma-$finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For example when F is the distribution function of a probability measure, then this measure is denoted $\mathbf{F}$ or dF. Denote in this case

$$F^{-1}(u) := \inf\{x \in \mathbb{R} \text{ s.t. } F(x) \geq u\} \text{ for } u \in (0,1)$$

the generalized inverse of F , a left continuous non decreasing function which is the quantile function of the probability measure $\mathbf{F}$. Denote accordingly $\mathbf{F}^{-1}$ or dF^{-1}, indifferently, the quantile measure with distribution function F^{-1}. If $x_1, .., x_n$ are n realizations of a random variable X with absolutely continuous probability measure $\mathbf{F}$ then the gaps in the empirical distribution function

$$F_n(x) := \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(x_i)$$

are of size $1/n$ and are located on the X_i's; the empirical quantile function satisfies

$$F_n^{-1}(u) = x_{i:n} \text{ when } \frac{i-1}{n} < u \leq \frac{i}{n}$$

and its gaps are given by

$$F_n^{-1}\left((i/n)^+\right) - F_n^{-1}\left((i/n)\right) = \mathbf{F}_n^{-1}\left(i/n\right) = x_{i+1:n} - x_{i:n}$$

where $x_{1:n} \leq ... \leq x_{n:n}$ denotes the ordered sample; those gaps will be denoted $\mathbf{F}_n^{-1}\left(i/n\right)$ or $dF_n^{-1}\left(i/n\right)$ indifferently; the empirical quantile measure has as its support the uniformely sparsed points $\{1/n, 2/n, .., 1\}$ and attributes masses equal to sampled spacings at those points; it follows that the empirical quantile measure is a positive finite measure with finite support. The quantile measure associated with the distribution function F^{-1} is also a positive $\sigma-$finite measure, defined on $(0,1)$. The above construction defined the quantile measure from the probability measure, but the reciprocal construction will be used, starting from a quantile measure, defining its distribution function, turning to its inverse to define a distribution of a probability measure, and then to the probability measure itself.

We now turn back to our topics.

Models defined as in the above examples extend the classical parametric ones, and are defined through some constraints on the form of the distributions. They can be paralleled with models defined through moments conditions defined as follows.

Let θ in Θ , an open subset of $\mathbb{R}^d$ and let $g : (x, \theta) \in \mathbb{R} \times \Theta \to \mathbb{R}^l$ be a l-valued function, each component of which is parametrized by $\theta \in \Theta \subset \mathbb{R}^d$. Define

$$M_\theta := \left\{ \mathbf{F} \text{ s.t. } \int_{\mathbb{R}} g(x,\theta)\mathbf{F}(dx) = 0 \right\}$$

and the semi parametric model defined by moment conditions is the collection of probability measures in

$$\mathcal{M} := \bigcup_{\theta \in \Theta} M_\theta. \tag{1.4}$$

These semiparametric models are defined by l conditions pertaining to l moments of the distributions and are widely used in applied statistics. When the dimension d of the parameter space exceeds l, no plug-in method can achieve any inference on θ; however, various techniques have been proposed in this case; see for example Hansen [62], who defined the Generalized Method of Moments (GMM) and Owen, who defined the so-called empirical likelihood approach [85]. Later, Newey and Smith [82] or Broniatowski and Keziou [29] proposed a refinement of the GMM approach minimizing a divergence criterion over the model. A major feature of models defined by (1.4) lies in their linearity with respect to the cumulative distribution function (cdf) which brings a dual formulation of the minimization problem. Duality results easily lead to the consistency and the asymptotic normality of the estimators of θ; see [29][82].

Similarly as for models defined by (1.4), we can introduce semiparametric linear quantile (SPLQ) models through

$$\bigcup_{\theta \in \Theta} L_\theta := \bigcup_{\theta \in \Theta} \left\{ \mathbf{F} \text{ s.t. } \int_0^1 F^{-1}(u)k(u,\theta)du = f(\theta) \right\} \tag{1.5}$$

where $\Theta \subset \mathbb{R}^d$, $k : (u, \theta) \in [0; 1] \times \Theta \to \mathbb{R}^l$ and $f : \Theta \to \mathbb{R}^l$. In the above display, in accordance with the above notation, F^{-1} denotes the generalized inverse function of F, the

distribution function of the measure $\mathbf{F}$. Examples 1.3,1.4 and 1.5 can be written through (1.5) ; see Section 1.3.2. We will consider the case when k is a function of u only ; this class contains many examples, typically models defined by a finite number of constraints on functions of the moments of the order statistics.
It is natural to propose similar estimation procedures for SPLQ models based on a minimization of a divergence. Models (1.5) do not enjoy linearity with respect to the cdf but with respect to the quantile function. Thus, as developed for models defined by (1.4), we propose to minimize a divergence criterion built on quantiles.
We will reformulate this criterion into a minimization of the energy of a deformation of the empirical distribution. In Section 1.6, a parallel with physical elasticity theory will be made as well as a comparison with an optimal transportation approach. We will prove a duality result and the subsequent consistency and asymptotic normality for the corresponding family of estimators. This will be done in Sections 1.5 and 1.7.

In the following, the transpose of a vector A will be expressed by A^T and if F and G are two cdf's, $F \ll G$ means that F is absolutely continuous with respect to G. The Lebesgue measure on $\mathbb{R}$ is denoted $d\lambda$ or dx, according to the common use in the context.

1.2 L-moments

1.2.1 Definition and characterizations

Let us consider data consisting in $\underline{X} = (x_1, ..., x_r)$, which are r realizations of real-valued independent and identically distributed (iid) copies $X_1, .., X_r$ of a random variable (r.v.) X with distribution function F. The r-th L-moment λ_r is defined as a particular L-statistics

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} \mathbb{E}[X_{r-k:r}] \tag{1.6}$$

where $X_{1:r} \leq X_{2:r} \leq ... \leq X_{r:r}$ denotes the order statistics of $X_1, .., X_r$.
From the above definition all L-moments λ_r but λ_1 are shift invariant, hence independent upon λ_1. If F is continuous, the expectation of the j-th order statistics $X_{j:r}$ is (see David p.33[45])

$$\mathbb{E}[X_{j:r}] = \frac{r!}{(j-1)!(r-j)!}\int_{\mathbb{R}} xF(x)^{j-1}(1-F(x))^{r-j}\mathbf{F}(dx). \tag{1.7}$$

The first four L-moments are

$$\begin{aligned}
\lambda_1 &= \mathbb{E}[X] \\
\lambda_2 &= \tfrac{1}{2}\mathbb{E}[X_{2:2} - X_{1:2}] \\
\lambda_3 &= \tfrac{1}{3}\mathbb{E}[X_{3:3} - 2X_{2:3} + X_{1:3}] \\
\lambda_4 &= \tfrac{1}{4}\mathbb{E}[X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}].
\end{aligned}$$

Remark 1.1. *The second L-moment is equal to the half of the absolute mean difference*

$$\lambda_2 = \frac{1}{2}\mathbb{E}[|X - Y|]$$

where X and Y are independently sampled from the same distribution F. The ratio $\frac{\lambda_2}{\lambda_1}$ is known as the Gini coefficient.

The expectations of the extreme order statistics characterize a distribution : if $\mathbb{E}(|X|)$ is finite, either of the sets $\{\mathbb{E}(X_{1:n}), n = 1, ..\}$ or $\{\mathbb{E}(X_{n:n}), n = 1, ..\}$ characterize the

distribution of X; see [37] and [71]. Since the moments of order statistics are defined by the family of L-moments, those also characterize the distribution of X.

The $r-$th L-moment ratio is defined for $r \geq 2$ by

$$\tau_r = \frac{\lambda_r}{\lambda_2}.$$

The interpretation of $\lambda_1, \lambda_2, \tau_3, \tau_4$ as measures of location, scale, skewness and kurtosis respectively and the existence of all L-moments whenever $\int |x|\mathbf{F}(dx) < \infty$ makes them good alternatives to moments.

Remark 1.2. *We can define from the quantile function $F^{-1} : [0;1] \to \mathbb{R}$ an associated measure on $\mathcal{B}([0;1])$*

$$\mathbf{F}^{-1}(B) = \int_0^1 \mathbb{1}_{x\in B} dF^{-1}(x) \in \mathbb{R} \cup \{-\infty, +\infty\}$$

The above integral is a Riemann-Stieltjes integral. It defines a σ-finite measure since F^{-1} has bounded variations on every interval of the form $[a,b]$ with $0 < a \leq b < 1$. For any $\mathbf{F}^{-1}$-measurable function $a : \mathbb{R} \to \mathbb{R}$, it holds

$$\int_0^1 a(x) dF^{-1}(x) = \int_0^1 a(x)\mathbf{F}^{-1}(dx)$$

Writing the L-moments of a distribution F as an inner product of the corresponding quantile function with a specific complete orthonormal system of polynomials in $L^2(0,1)$ is a cornerstone in the derivation of statistical inference in SPLQ models. The shifted Legendre polynomials define such a system of functions.

Definition 1.3. *The shifted Legendre polynomial of order r is*

$$L_r(t) = \sum_{k=0}^{r} (-1)^k \binom{r}{k}^2 t^{r-k}(1-t)^k = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k}\binom{r+k}{k} t^k. \tag{1.8}$$

Let us define for $r \geq 1$, K_r as the integrated shifted Legendre polynomials

$$K_r(t) = \int_0^t L_{r-1}(u)du = -t(1-t)\frac{J_{r-2}^{(1,1)}(2t-1)}{r-1} \tag{1.9}$$

with $J_{r-2}^{(1,1)}$ the corresponding Jacobi polynomial (see [63])

$$J_{r-2}^{(1,1)}(2t-1) = \frac{\Gamma(r)}{(r-2)!\Gamma(r+1)} \sum_{k=0}^{r-2} \binom{k}{r-2} \frac{\Gamma(r+1+k)}{\Gamma(2+k)} (t-1)^k.$$

We can state the following result.

Proposition 1.1. *Let F be any cdf and assume that $\int |x|\, dF(x)$ is finite. Then for any $r \geq 1$, it holds*

$$\lambda_r = \int_0^1 F^{-1}(t) L_{r-1}(t) dt = \int_0^1 F^{-1}(t) dK_r(t) \tag{1.10}$$

where the last integral is the Stieltjes integral of F^{-1} with respect to the function $t \mapsto K_r(t)$.

Proof. The proof is based on the following fundamental Lemma, whose proof is deferred to the Appendix.

Lemma 1.1. *Let U be a uniform random variable on [0;1] and X the random variable associated to the cdf F. Then $F^{-1}(U) =_d X$.*

Let $U_1, ..., U_r$ be r independent random variable uniformly distributed on $[0;1]$ and denote by $U_{1:r} \leq ... \leq U_{r:r}$ r the ordered statistics. Then

$$(X_{1:r}, ..., X_{r:r}) \stackrel{d}{=} (F^{-1}(U_{1:r}), ..., F^{-1}(U_{r:r}));$$

hence for $1 \leq j \leq r$

$$\mathbb{E}[X_{j:r}] = \mathbb{E}[F^{-1}(U_{j:r})] = \frac{r!}{(j-1)!(r-j)!} \int_0^1 F^{-1}(t) t^{j-1}(1-t)^{r-j} dt,$$

which ends the proof. □

Before going any further, we present an useful Lemma, the proof of which is also deferred to the Appendix.

Lemma 1.2. *Let a be a real-valued function such that $\int_{\mathbb{R}} a(x)dF(x) < \infty$. Then*

$$\int_{\mathbb{R}} a(x) d\mathbf{F}(x) = \int_0^1 a(F^{-1}(t))dt. \tag{1.11}$$

Similarly if $t \to b(t)$ is a real-valued function such that $\int_0^1 b(t)\mathbf{F}^{-1}(dt) < \infty$. Then

$$\int_0^1 b(t)\mathbf{F}^{-1}(dt) = \int_0^1 b(F(x))dx. \tag{1.12}$$

Remark 1.3. *As a consequence of Lemma 1.2 and equation (1.10), it holds*

$$\lambda_r = \int_0^1 x dK_r(F(x)).$$

Remark 1.4. *If we consider a multinomial distribution with support $x_1 \leq x_2 \leq ... \leq x_n$ and associated weights $\pi_1, ..., \pi_n$ ($\sum_{i=1}^n \pi_i = 1$), we get*

$$\lambda_r = \sum_{i=1}^n w_i^{(r)} x_i = \sum_{i=1}^n \left[K_r\left(\sum_{a=1}^i \pi_a\right) - K_r\left(\sum_{a=1}^{i-1} \pi_a\right)\right] x_i = \int_0^1 L_{r-1}(t) Q_\pi(t) dt$$

with

$$Q_\pi(t) = \begin{cases} x_1 & \text{if} \quad 0 \leq t \leq \pi_1 \\ x_i & \text{if} \quad \sum_{a=1}^{i-1} \pi_a < t \leq \sum_{a=1}^{i} \pi_a \end{cases}.$$

This example illustrates Remark 1.3.

Figure 1.1 provides the first weight $w_i^{(r)}$ when the x_i's are equally sparsed on $[0,1]$ with equal weights $\pi_1 = .. = \pi_n = 1/n$.

The following characterization for the L-moments with order larger or equal to 2 is used in Section 1.3.2.

Proposition 1.2. *If $r \geq 2$ and $\int_{\mathbb{R}} |x| \, dF(x) < +\infty$, then*

$$\lambda_r = \int_0^1 F^{-1}(t) dK_r(t) = -\int_0^1 K_r(t) \mathbf{F}^{-1}(dt). \tag{1.13}$$

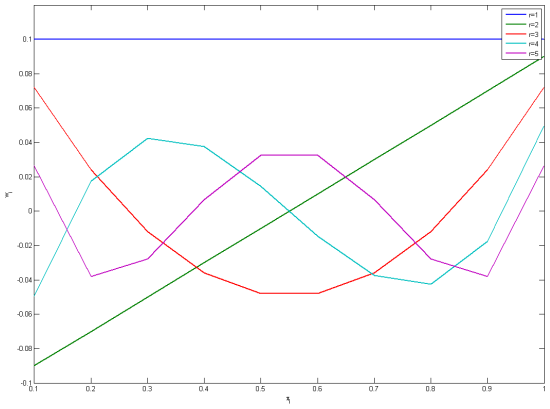


FIGURE 1.1 – Weights $w_i^{(r)}$ for the uniform law with a support containing 10 points

Proof. This result follows as an application of the Fubini-Tonelli theorem. Indeed

$$\begin{aligned}\lambda_r &= \int_0^1 F^{-1}(t)dK_r(t) \\ &= \int_0^1 \int_0^t \mathbf{F}^{-1}(du)dK_r(t) \\ &= \int_0^1 \int_0^1 \mathbb{1}_{0\leq u\leq t}\mathbf{F}^{-1}(du)dK_r(t).\end{aligned}$$

This last equality holds since $(u,t) \mapsto \mathbb{1}_{0\leq u\leq t}$ is measurable with respect to the measure $\mathbf{F}^{-1} \times dK_r$ since $\mathbb{E}[X] < \infty$. Applying Fubini-Tonelli Theorem, it holds

$$\begin{aligned}\lambda_r &= \int_0^1 \int_0^1 \mathbb{1}_{0\leq u\leq t}dK_r(t)\mathbf{F}^{-1}(du) \\ &= \int_0^1 \int_0^1 \left[K_r(1) - K_r(u)\right]\mathbf{F}^{-1}(du) \\ &= -\int_0^1 K_r(u)\mathbf{F}^{-1}(du)\end{aligned}$$

since $K_r(1) = 0$ for $r > 1$. □

Remark 1.5. *That (1.13) does not hold for $r = 1$ follows from the fact that if $G = F(.+a)$ for some $a \in \mathbb{R}$, then $\mathbf{G}^{-1} = \mathbf{F}^{-1}$. Hence, SPLQ models are shift-invariant. This can also be seen setting $r = 1$ in the right-hand side of (1.13); in this case, the integral is infinite (but if supp$(\mathbf{F})$ is bounded) whereas λ_1 is supposed to be finite.*

1.2.2 Estimation of L-moments

Let $x_1, ..., x_n$ be iid realizations of a random variable X with distribution F and L-moments λ_r. Define F_n the empirical cdf of the sample and l_r the corresponding plug-in estimator of λ_r,

$$l_r = \int_0^1 F_n^{-1}(t)L_{r-1}(t)dt. \tag{1.14}$$

This estimator of λ_r is biased as quoted in [64] and [99]. l_r is usually termed as a V-statistic. As noted upon in [64] and [99], the unbiased estimators of L-moments are the following U-statistics

$$l_r^{(u)} = \frac{1}{\binom{n}{r}} \sum_{1 \le i_1 < \cdots < i_r \le n} \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} x_{i_{r-k}:n}.$$

Remark 1.6. *An alternative definition for l_r as in (1.14) can be stated as follows. Conditionally on the realizations $x = (x_1, ..., x_n)$, define the uniform distribution on x. Then l_r is the discrete L-moment of order r of this conditional distribution. It can therefore be defined through*

$$l_r = \frac{1}{\binom{r+n-1}{n-1}} \sum_{1 \le i_1 \le \cdots \le i_r \le n} \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} x_{i_{r-k}:n}.$$

Let us now extend Definition 1.6 of the L-moments as follows. Let $(i_1, ..., i_r)$ be drawn without replacement from $\{1, ..., r\}$. We then define $x_{(i_1)} \le ... \le x_{(i_r)}$ the corresponding ordered observations and

$$\lambda_r^{(u)} = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} \mathbb{E}[x_{(i_{r-k})}]$$

where the expectation is taken under the extraction process. Then $\lambda_r^{(u)}$ and $l_r^{(u)}$ coincide. Although $l_r^{(u)}$ is unbiased, for sake of simplicity only l_r which is asymptotically unbiased, will be used in the sequel.

These two estimators l_r and $l_r^{(u)}$ of the L-moment λ_r have the same asymptotic properties.

Proposition 1.3. *Let us suppose that F has finite variance. Then, for any $m \ge 1$*

$$\sqrt{n} \left[\begin{pmatrix} l_1 \\ \vdots \\ l_m \end{pmatrix} - \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_m \end{pmatrix} \right] \to_d \mathcal{N}_m(0, \Lambda)$$

where $\mathcal{N}_m$ denotes the multivariate normal distribution and the elements of Λ are given by

$$\Lambda_{rs} = \int \int_{x<y} \left[L_{r-1}(F(x)) L_{s-1}(F(y)) + L_{r-1}(F(y)) L_{s-1}(F(x)) \right] F(x)(1 - F(y)) dx dy$$

Furthermore, the same property holds for $l_1, .., l_r$ substituted by $l_1^{(u)}, ..., l_m^{(u)}$.

Proof. This is a plain consequence of Theorem 6 in [100]. See also [64] for an evaluation of the bias of l_r. □

1.3 Models defined by moment and L-moment equations

1.3.1 Models defined by moment conditions

Let us consider n iid random variables X_1,...,X_n drawn from the same distribution function F. Semi-parametric models are often defined through equations :

$$\int_{\mathbb{R}} g(x,\theta)\mathbf{F}(dx) = \mathbb{E}[g(X,\theta)] = 0$$

where $g : \mathbb{R} \times \Theta \to \mathbb{R}^l$ and $\Theta \subset \mathbb{R}^d$ is a space of parameters, as quoted in Section 1.1.

Example 1.6. *We can sometimes face distributions with constraints pertaining to the two first moments. For example, Godambe and Thompson [56] considered the distributions verifying* $\mathbb{E}[X] = \theta$ *and* $\mathbb{E}[X^2] = h(\theta)$ *with a known function* h*. Then, with our notations* $l = 2$ *and* $g(x,\theta) = (x - \theta, x^2 - h(\theta))$

Example 1.7. *Consider the distributions* F *such that for some* θ *it holds* $F(y) = 1 - F(-y) = \theta$ *[29]. This corresponds to a moment condition model with* $l = 2$ *and* $g(x,\theta) = (\mathbb{1}_{]-\infty;y]}(x) - \theta, \mathbb{1}_{[y;+\infty[}(x) - \theta)$*. The condition on the model is the existence of some* θ *such that the left and right quantiles of order* θ *are* $-y$ *and* $+y$ *for some given* y*.*

1.3.2 Models defined by L-moments conditions

In the present paper we consider models defined by l constraints on their first L-moments, namely satisfying

$$-\mathbb{E}\left[\frac{1}{r}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}X_{k:r}\right] = f_r(\theta) \quad 1 \le r \le l \tag{1.15}$$

where Θ is some open set in $\mathbb{R}^d$ and $f_j : \Theta \to \mathbb{R}$ are some given functions defined on Θ. Those models are SPLQ, with $(u,\theta) \mapsto k(u,\theta)$ independent on θ, defined by

$$k(u,\theta) = -L(u) := -\begin{pmatrix} L_1(u) \\ \vdots \\ L_l(u) \end{pmatrix} \tag{1.16}$$

where the shifted Legendre polynomials L_j have been defined in Definition 1.3.

The SPLQ model (1.5) may be written as

$$\mathcal{L} := \bigcup_{\theta\in\Theta} L_\theta = \bigcup_{\theta\in\Theta}\left\{\mathbf{F} \text{ s.t. } \int_0^1 L(u)F^{-1}(u)du = -f(\theta)\right\}. \tag{1.17}$$

Due to Proposition 1.2 we may write equation (1.15) for $r \ge 2$ as follows, making use of the integrated shifted Legendre polynomials K_r in lieu of L_r.

$$-\mathbb{E}\left[\frac{1}{r}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}X_{k:n}\right] = \int_0^1 K_r(u)\mathbf{F}^{-1}(du) = f_r(\theta). \tag{1.18}$$

Example 1.8. *Turning back to Example 1.3, we define k and f by*

$$k(u,\theta) = -\begin{pmatrix} L_2(u) \\ L_3(u) \\ L_4(u) \end{pmatrix}$$

and

$$f(\theta) = \begin{pmatrix} f_2(\theta) \\ f_3(\theta) \\ f_4(\theta) \end{pmatrix} = \begin{pmatrix} \sigma(1-2^{-1/\nu})\Gamma(1+1/\nu) \\ f_2(\theta)[3-2\frac{1-3^{-1/\nu}}{1-2^{-1/\nu}})] \\ f_2(\theta)[6+\frac{5(1-4^{-1/\nu})-10(1-3^{-1/\nu})}{1-2^{-1/\nu}})] \end{pmatrix}$$

where $\theta = (\sigma,\nu) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$ and $u \in [0;1]$.

Example 1.9. *Similarly, in case we consider Example 1.4, we define k and f by*

$$k(u,\theta) = -\begin{pmatrix} L_2(u) \\ L_3(u) \\ L_4(u) \end{pmatrix}$$

and

$$f(\theta) = \begin{pmatrix} f_2(\theta) \\ f_3(\theta) \\ f_4(\theta) \end{pmatrix} = \begin{pmatrix} \frac{\sigma}{(1+\nu)(2+\nu)} \\ f_2(\theta)\frac{1-\nu}{3+\nu} \\ f_2(\theta)\frac{(1-\nu)(2-\nu)}{(3+\nu)(4+\nu)} \end{pmatrix}$$

where $\theta = (\sigma,\nu) \in \mathbb{R}_+^* \times \mathbb{R}$ and $u \in [0;1]$.

1.3.3 Extension to models defined by order statistics conditions

The order statistics given by equation (1.7) can be written as

$$\mathbb{E}[X_{j:r}] = \int_0^1 P_{j:r}(u)F^{-1}(u)du$$

where the polynomials $P_{j:r}$ are given by

$$P_{j:r}(u) = \frac{r!}{(j-1)!(r-j)!}u^{j-1}(1-u)^{r-j}.$$

Any L-statistics can therefore be written

$$-\sum_{i=1}^{r} a_j \mathbb{E}[X_{j:r}] = \int_0^1 P_a(u)F^{-1}(u)du$$

with coefficients a_j's belonging to $\mathbb{R}$ and

$$P_a(u) = -\sum_{i=1}^{r} a_j P_{j:r}(u).$$

These models are SPLQ (see 1.5) with

$$\mathcal{L} := \bigcup_\theta L_\theta = \bigcup_\theta \left\{ F \text{ s.t. } \int_0^1 P(u)F^{-1}(u)du = -f(\theta) \right\} \tag{1.19}$$

where $P : u \in [0;1] \mapsto P(u) \in \mathbb{R}^l$ is an array of l polynomials.

Example 1.10. *Turning back to Example 1.5, we define k and f by*

$$k(u,\theta) = \begin{pmatrix} P_{1:3}(u) \\ P_{2:3}(u) \\ P_{3:3}(u) \end{pmatrix}$$

and

$$f(\theta) = \begin{pmatrix} \theta - \nu \\ \theta \\ \theta + \nu \end{pmatrix}$$

where $\theta \in \mathbb{R}, \nu > 0$ and $u \in [0;1]$.

1.4 Minimum of φ-divergence estimators

Estimation, confidence regions and tests based on moment conditions models have evolved over thirty years. Hansen and Owen respectively proposed the generalized method of moments (GMM)[61] and the empirical likelihood (EL) estimators [85]. Newey and Smith [82] introduced the generalized empirical likelihood (GEL) family of estimators encompassing the previous estimators. They also proposed the dual versions of the GEL estimators, the minimum discrepancy estimators (MD). These estimators are the solution of the minimization of a divergence with constraints corresponding to the model; see also Broniatowski and Keziou [29] for an approach through duality and properties of the inference under misspecification. In the quantiles framework, Gourieroux proposed an adaptation of GMM estimators in [57] for a parametric model seen through its quantile function $F^{-1}(t,\theta)$. In the following, we will consider inference based on divergences in order to present estimators for models defined by L-moments conditions.

1.4.1 φ-divergences

Let $\varphi : \mathbb{R} \to [0, +\infty]$ be a strictly convex function with $\varphi(1) = 0$ such that $\text{dom}(\varphi) = \{x \in \mathbb{R} | \varphi(x) < \infty\} := (a_\varphi, b_\varphi)$ with $a_\varphi < 1 < b_\varphi$. If F and G are two σ-finite measures of $(\mathbb{R}, B(\mathbb{R}))$ such that G is absolutely continuous with respect to F, we define the divergence between F and G by :

$$D_\varphi(G,F) = \int_{\mathbb{R}} \varphi\left(\frac{dG}{dF}(x)\right) dF(x) \tag{1.20}$$

where $\frac{dG}{dF}$ is the Radon-Nikodym derivative. It is clear that when $F = G$, $D_\varphi(F,G) = 0$. Furthermore, as φ is supposed to be strictly convex,

$$D_\varphi(G,F) = 0 \text{ if and only if } F = G.$$

These divergences were independently introduced by Csiszar [42] or Ali and Silvey [2] in the context of probability measures. Note that we gave Definition 1.20 for any σ-finite measures even if our notations refer to probability measures. Indeed, we will consider in the sequel divergence between quantile measure which are σ-finite but may be not finite. See Liese [76] who also considered divergences between σ-finite measures.

Example 1.11. *The class of power divergences parametrized by $\gamma \geq 0$ is defined through the functions*

$$x \mapsto \varphi_\gamma(x) = \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma - 1)}.$$

The domain of φ_γ depends on γ. The Kullback-Leibler divergence is associated to $x > 0 \mapsto \varphi_1(x) = x\log(x) - x + 1$, the modified Kullback-Leibler (KL_m) divergence to $x > 0 \mapsto \varphi_0(x) = -\log(x) + x - 1$, the χ^2-divergence to $x \in \mathbb{R} \mapsto \varphi_2(x) = 1/2(x-1)^2$, etc.

1.4.2 M-estimates with L-moments constraints

Minimum of φ-divergences for probability measures

A plain approach to inference on θ consists in mimicking the empirical minimum divergence one, substituting the linear constraints with respect to the distribution by the corresponding linear constraints with respect to the quantile measure, and minimizing the divergence between all probability measures satisfying the constraint and the empirical measure F_n pertaining to the data set. More formally this yields to the following program.

Denote by M the set of all probability measures defined on $\mathbb{R}$. For a given p.m. $\mathbf{F}$ in M we consider the submodel which consists in all p.m's $\mathbf{G}$ in M, absolutely continuous with respect to F, and which satisfy the constraints on their first L-moments for a given $\theta \in \Theta$. Identifying a measure $\mathbf{G}$ with its distribution function G we define

$$L_\theta^{(0)}(\mathbf{F}) = \{\mathbf{G} \in M \text{ s.t. } \mathbf{G} \ll \mathbf{F}, \int_0^1 L(t)G^{-1}(t)dt = -f(\theta)\}.$$

Probability measures $\mathbf{G}$ satisfying the constraints and bearing their mass on the sample points belong to $L_\theta^{(0)}(\mathbf{F}_n)$. For any parameter $\theta \in \Theta$, the distance between $\mathbf{F}$ and the submodel $L_\theta^{(0)}(\mathbf{F})$ is defined by

$$D_\varphi(L_\theta^{(0)}(\mathbf{F}), \mathbf{F}) = \inf_{\mathbf{G} \in L_\theta^{(0)}(\mathbf{F})} D_\varphi(\mathbf{G}, \mathbf{F}),$$

and its plug-in estimator is

$$D_\varphi(L_\theta^{(0)}(\mathbf{F}_n), \mathbf{F}_n) = \inf_{\mathbf{G} \in L_\theta^{(0)}(\mathbf{F}_n)} D_\varphi(\mathbf{G}, \mathbf{F}_n).$$

which measures the distance between the empirical measure $\mathbf{F}_n$ and the class of all the probability measures supported by the sample and which satisfy the L-moment conditions for a given θ.

A natural estimator for θ may be defined by :

$$\hat{\theta}_n^{(0)} = \arg\inf_{\theta \in \Theta} D_\varphi(L_\theta^{(0)}(\mathbf{F}_n), \mathbf{F}_n) = \arg\inf_{\theta \in \Theta} \inf_{\mathbf{G} \in L_\theta^{(0)}(F_n)} \frac{1}{n}\sum_{i=1}^n \varphi(n\mathbf{G}(x_i)). \tag{1.21}$$

Unfortunately, existence of this estimator may not hold. Indeed, we cannot assess that $L_\theta^{(0)}(\mathbf{F}_n)$ is not empty : its elements are solutions of a polynomial algebraic equation of degree l. To our knowledge, general conditions of existence for the solutions of such problems do not exist.

Bertail in [19] proposes a linearization of the constraint in (1.21). We here prefer to switch to a different approach. If we consider the L-moment equation (1.18), we see that the quantile function plays a similar role as the distribution function in the classical moment equations. We will then change the functional to be minimized in order to be able to use duality for the optimization step.

Minimum of φ-divergences for quantile measures

We have seen that the characterization of the L-moments given by the equation (1.18) uses the quantile measure $\mathbf{F}^{-1}$, which is defined by the generalized inverse function of F. If $\mathbf{F}^{-1}$ is absolutely continuous, we can define the quantile-density $q(u) = (F^{-1})'(u)$. This density was called "sparsity" function by Tukey [103] as it represents the sparsity of the distribution at the cumulating weight $u \in [0;1]$. This is clear when we look at the empirical version of this measure which is composed by nothing but the increments of the sample. Some other approach, handling properties of the inverse function of $(F^{-1})'$, have been proposed by Parzen [87]. He claims that the inference procedures based on $(F^{-1})'$ possesses inherent robustness properties.

Define

$$K(u) = \begin{pmatrix} K_2(u) \\ \vdots \\ K_l(u) \end{pmatrix}$$

and

$$f^{(2:l-1)}(u) = \begin{pmatrix} f_2(u) \\ \vdots \\ f_l(u) \end{pmatrix}.$$

For any θ in Θ the submodel which consists of all p.m's $\mathbf{G}$ with mass on the sample points is substituted by the set of all quantile measures denoted $\mathbf{G}^{-1}$ which have masses on subsets of $\{1/n, 2/n, .., 1\}$ and whose distribution functions coincide with the generalized inverse functions of elements in $L_\theta^{(0)}(\mathbf{F}_n)$.
Let N be the class of all $\sigma-$finite positive measures on $\mathbb{R}$. Making use of equation (1.18) define

$$\begin{aligned} L_\theta^n &:= \left\{ \mathbf{G}^{-1} \in N \text{ s.t. } \mathbf{G}^{-1} \ll \mathbf{F}_n^{-1} \text{ and } \int_0^1 L(u) G^{-1}(u) du = -f(\theta) \right\} \\ &= \left\{ \mathbf{G}^{-1} \in N \text{ s.t. } \mathbf{G}^{-1} \ll \mathbf{F}_n^{-1} \text{ and } \int_0^1 K(u) \mathbf{G}^{-1}(du) = f^{(2:l-1)}(\theta) \right\} \end{aligned}$$

the family of all measures $\mathbf{G}^{-1}$ with support included in $\{1/n, 2/n, .., 1\}$ which satisfy the $l-1$ constraints pertaining to the L-moments ; see (1.17). Note that when $\mathbf{F}$ bears an atom then for large enough n then $\mathbf{G}^{-1}$ in L_θ^n has a support strictly included in $\{1/n, 2/n, .., 1\}$.

A natural proposal for an estimation procedure in the SPLQ model is then to consider the minimum of a φ-divergence between quantile measures through

$$\hat{\theta}_n = \arg \inf_{\theta \in \Theta} \inf_{\mathbf{G}^{-1} \in L_\theta^n} \int_0^1 \varphi \left(\frac{d\mathbf{G}^{-1}}{d\mathbf{F}_n^{-1}}(u) \right) \mathbf{F}_n^{-1}(du) \tag{1.22}$$

$$= \arg \inf_{\theta \in \Theta} \inf_{y := (y_1 \le \ldots \le y_n); \sum_{i=1}^{n-1} K(i/n)(y_{i+1} - y_i) = f(\theta)} \sum_{i=1}^{n-1} \varphi \left(\frac{y_{i+1} - y_i}{x_{i+1:n} - x_{i:n}} \right) (x_{i+1:n} - x_{i:n}). \tag{1.23}$$

Remark 1.7. *The estimation defined by (1.22) produces estimators $\hat{\theta}_n$ which do not depend on the location of the sample, since a change the sample $(x_i \mapsto x_i + a)_{i=1\ldots n}$ produces, independently on the value of a, the same measure $\mathbf{F}_n^{-1}$ whose mass on point i/n is the gap $x_{i+1:n} - x_{i:n}$. The minimum discrepancy estimators defined by (1.23) are invariant with respect to the location of the underlying distribution of the data. Due to this fact, we*

consider the model defined by L-moments conditions only through equations of the form (1.18).

Both the constraint and the divergence criterion are expressed in function of $\mathbf{G}^{-1}$ and the constraint is linear with respect to this measure. This allows to use classical duality results in order to efficiently compute the estimator $\hat{\theta}_n$. Before that, we reformulate this criterion as a minimization of an "energy" of transformation of the sample.

1.5 Dual representations of the divergence under L-moment constraints

The minimization of φ-divergences under linear equality constraint is performed using Fenchel-Legendre duality. It transforms the constrained problems into an unconstrained one in the space of Lagrangian parameters. Let ψ denote the Fenchel-Legendre transform of φ, namely, for any $t \in \mathbb{R}$

$$\psi(t) := \sup_{x \in \mathbb{R}} \{tx - \varphi(x)\}.$$

Let us recall that $dom(\varphi) = (a_\varphi, b_\varphi)$. We can now present a general duality result for the two optimization problems that transform a constrained problem (possibly in an infinite dimensional space) into an unconstrained one in $\mathbb{R}^l$.

Let $C : \Omega \to \mathbb{R}^l$ and $a \in \mathbb{R}^l$. Denote

$$L_{C,a} = \left\{ g : \Omega \to \mathbb{R} \text{ s.t. } \int_\Omega g(t)C(t)\mu(dt) = a \right\}.$$

Proposition 1.4. *Let μ be a σ-finite measure on $\Omega \subset \mathbb{R}$. Let $C : \Omega \to \mathbb{R}^l$ be an array of functions such that*

$$\int_\Omega \|C(t)\|\mu(dt) < \infty.$$

If there exists some g in $L_{C,a}$ such that $a_\varphi < g < b_\varphi$ μ-a.s. then the duality gap is zero i.e.

$$\inf_{g \in L_{C,a}} \int_\Omega \varphi(g)\, d\mu = \sup_{\xi \in \mathbb{R}^l} \langle \xi, a \rangle - \int_\Omega \psi(\langle \xi, C(x) \rangle)\mu(dx). \tag{1.24}$$

Moreover, if ψ is differentiable, if μ is positive and if there exists a solution ξ^* of the dual problem which is an interior point of

$$\left\{ \xi \in \mathbb{R}^l \text{ s.t. } \int_\Omega \psi(\langle \xi, C(x) \rangle)\mu(dx) < \infty \right\},$$

then ξ^* is the unique maximum in (1.24) and

$$\int \psi'(\langle \xi^*, C(x) \rangle)C(x)\mu(dx) = a.$$

Furthermore the mapping $a \mapsto \xi^*(a)$ is continuous.

Proof. The proof is delayed to the Appendix. □

Remark 1.8. *When $\mathbf{G}^{-1} \ll \mathbf{F}^{-1}$, denoting $g^* = d\mathbf{G}^{-1}/d\mathbf{F}^{-1}$ and assuming $g^* \in L_{K,f(\theta)}$, and when $\mu = \mathbf{F}^{-1}$ it holds*

$$\int \varphi(g^*)d\mu = D_\varphi\left(\mathbf{G}^{-1}, \mathbf{F}^{-1}\right).$$

Remark 1.9. *Here, the classical assumption of finiteness of μ is replaced by*

$$\int_\Omega \|C(x)\|\mu(dx) < \infty$$

which is needed for the application of the dominated convergence Theorem ; also we refer to the illuminating paper by Csiszár and Matúš [44] for the description of the geometric tools used in the proof of Proposition 1.4.

We now apply the above Proposition 1.4 to the case when the array of functions C is equal to K, the measure μ is the quantile measure $\mathbf{F}^{-1}$ pertaining to the distribution function F of a probability measure and when the class of functions $L_{C,a}$ is substituted by the class of functions $\mathbf{dG}^{-1}/\mathbf{dF}^{-1}$ when defined. Let $\theta \in \Theta$ and F be fixed. Let us recall that for any reference cdf F

$$L_\theta(\mathbf{F}^{-1}) := \left\{ \mathbf{G}^{-1} \ll \mathbf{F}^{-1} \text{ s.t. } \int_{\mathbb{R}} K(u)\mathbf{G}^{-1}(du) = f(\theta) \right\}. \tag{1.25}$$

Corollary 1.1. *If there exists some $\mathbf{G}^{-1}$ in $L_\theta(\mathbf{F}^{-1})$ such that $a_\varphi < d\mathbf{G}^{-1}/d\mathbf{F}^{-1} < b_\varphi$ $\mathbf{F}^{-1}$-a.s. then*

$$\inf_{\mathbf{G}^{-1}\in L_\theta(\mathbf{F}^{-1})} \int_0^1 \varphi\left(\frac{d\mathbf{G}^{-1}}{d\mathbf{F}^{-1}}\right) d\mathbf{F}^{-1} = \sup_{\xi\in\mathbb{R}^l} \langle \xi, f(\theta)\rangle - \int_0^1 \psi(\langle \xi, K(u)\rangle)\mathbf{F}^{-1}(du). \tag{1.26}$$

Moreover, if ψ is differentiable and if there exists a solution ξ^* of the dual problem which is an interior point of

$$\left\{ \xi \in \mathbb{R}^l \text{ s.t. } \int_{\mathbb{R}} \psi(\langle \xi, K(u)\rangle)\mathbf{F}^{-1}(du) < \infty \right\},$$

then ξ^* is the unique maximum in (1.26) and

$$\int \psi'^*(\langle \xi, K(u)\rangle)K(u)\mathbf{F}^{-1}(du) = f(\theta).$$

Remark 1.10. *The above Corollary 1.1 is the cornerstone for the plug-in estimator of $D_\varphi(\mathbf{G}, \mathbf{F})$.*

Remark 1.11. *The model defined for the empirical quantile measure $L_\theta(\mathbf{F}^{-1})$ is equal to L_θ^n defned by equation 1.23.*

Let us present an other application of the above Proposition 1.4 leading to the same dual problem. Denote by λ the Lebesgue measure on $\mathbb{R}$ and $L'_\theta(F)$ be the set of all functions g defined by

$$L'_\theta(F) = \left\{ g : \mathbb{R} \to \mathbb{R} \text{ s.t. } \int_{\mathbb{R}} K(F(x))g(x)\lambda(dx) = f(\theta) \right\},$$

whenever non void.

Corollary 1.2. *If there exists some g in $L'_\theta(F)$ such that $a_\varphi < g < b_\varphi$ λ-a.s. then*

$$\inf_{g\in L'_\theta(F)} \int_{\mathbb{R}} \varphi(g)\, d\lambda = \sup_{\xi\in\mathbb{R}^l} \langle \xi, f(\theta)\rangle - \int_{\mathbb{R}} \psi(\langle \xi, K(F(x))\rangle)dx. \tag{1.27}$$

Moreover, if ψ is differentiable and if there exists a solution ξ^* of the dual problem which is an interior point of

$$\left\{\xi \in \mathbb{R}^l \ s.t. \ \int_{\mathbb{R}} \psi(\langle \xi, K(F(x))\rangle)dx < \infty\right\},$$

then ξ^* is the unique maximizer in (1.27). It satisfies

$$\int \psi'(\langle \xi^*, K(F(x))\rangle)dx = f(\theta) \tag{1.28}$$

Proof. We will detail the proof of Corollary 1.2. Corollary 1.1 is proved similarly.
We apply the above Proposition 1.4 for $\Omega = \mathbb{R}$, $\mu = \lambda$, the array of functions C substituted by the array of functions $x \mapsto K(F(x))$ and $a = f(\theta)$.
Consequently, the class of functions g, namely $L_{C,a}$, depends upon F. Following the notation of the Corollary

$$L'_\theta(F) = \left\{g : \mathbb{R} \to \mathbb{R}; \int_{\mathbb{R}} K(F(x))g(x)\lambda(dx) = f(\theta)\right\}.$$

We need then to show that

$$\int_{\mathbb{R}} \|K(F(x))\|dx < \infty.$$

Let us note $K = (K_{i_1}, ..., K_{i_l})$ with $i_j \geq 2$ for all j. Let then recall that from equation (1.9)

$$K_{i_j}(t) = -t(1-t)\frac{J^{(1,1)}_{i_j-2}(2t-1)}{i_j - 1}$$

It is clear that there exists $C > 0$ such that $\left|\frac{J^{(1,1)}_{i_j-2}(2t-1)}{i_j-1}\right| < C$. Hence

$$\int_{\mathbb{R}} \|K(F(x))\|dx < lC \int_{\mathbb{R}} F(x)(1-F(x))dx < +\infty$$

since F is the cdf of a random variable with finite expectation. By applying Proposition 1.4, it then holds

$$\inf_{g \in L''_\theta(F)} \int_{\mathbb{R}} \varphi(g)\, d\lambda = \sup_{\xi \in \mathbb{R}^l} \langle \xi, f(\theta)\rangle - \int_{\mathbb{R}} \psi(\langle \xi, K(F(x))\rangle)dx.$$

□

Remark 1.12. *If we consider the class of functions $T := x \mapsto \int_{-\infty}^x g(y)\lambda(dy)$, which is equal to*

$$L''_\theta(F) = \left\{T : \mathbb{R} \to \mathbb{R} \ s.t. \ T \ derivable \ \lambda - a.e. \ and \ \int_{\mathbb{R}} K(F(x))\frac{dT}{d\lambda}(x)\lambda(dx) = f(\theta)\right\},$$

rather than the class of functions g, it holds that $T \in L''_\theta(F)$ if and only if $dT/d\lambda \in L'_\theta(F)$. Therefore,

$$\inf_{T \in L''_\theta(F)} \int_{\mathbb{R}} \varphi\left(\frac{dT}{d\lambda}\right) d\lambda = \inf_{g \in L'_\theta(F)} \int_{\mathbb{R}} \varphi(g)\, d\lambda,$$

This seemingly formal definition of the function T makes sense since we can view T as a deformation function, as detailed in the following Section 1.6.

1.6 Reformulation of divergence projections and extensions

1.6.1 Minimum of an energy of deformation

The case of models defined by moments constraints

Let us suppose for a while that $\mathbf{F}$ and $\mathbf{G}$ are both absolutely continuous with respect to the Lebesgue measure defined on $\mathbb{R}$. Define the function $T = G \circ F^{-1}$. Then T is derivable a.e. and $T' = \frac{dT}{d\lambda}$. It holds

$$D_\varphi(\mathbf{G}, \mathbf{F}) = \int_{\mathbb{R}} \varphi\left(\frac{d\mathbf{G}}{d\mathbf{F}}(x)\right) \mathbf{F}(dx) = \int_0^1 \varphi\left(T'(u)\right) du$$

even if $\mathbf{G}$ is not a positive measure, as far as the integrand in the central term of the above display is defined.
The function T can be viewed as a measure of the deformation of $\mathbf{F}$ into $\mathbf{G}$ and

$$E_1(T) = \int \varphi\left(\frac{dT}{d\lambda}\right) d\lambda$$

as an energy of this deformation.

It can be seen that the absolute continuity assumption above can be relaxed.

Proposition 1.5. *Let F and G be two arbitrary cdf's and λ be the Lebesgue measure. Let us define*

$$M_\theta(\mathbf{F}) = \left\{\mathbf{G} \ll \mathbf{F} \ s.t. \ \int_{\mathbb{R}} g(x,\theta)\mathbf{G}(dx) = 0\right\}$$

and let $M'_\theta(\mathbf{F})$ denote the class of all functions T which are a.e derivable on $[0;1]$ defined through

$$M'_\theta(\mathbf{F}) = \left\{T : [0;1] \to \mathbb{R} \ s.t. \ a_\varphi < \frac{dT}{d\lambda} < b_\varphi \ \lambda - a.e. \ and \int_0^1 g(F^{-1}(u),\theta)\frac{dT}{d\lambda}(u)\lambda(du) = 0\right\}. \tag{1.29}$$

Then

$$\inf_{G \in M_\theta(\mathbf{F})} \int_{\mathbb{R}} \varphi\left(\frac{d\mathbf{G}}{d\mathbf{F}}(x)\right) \mathbf{F}(dx) = \inf_{T \in M'_\theta(\mathbf{F})} E_1(T).$$

Proof. This results from Proposition 1.4 applied twice.
First, if $C = g(.,\theta)$, $a = 0$, $\mu = \mathbf{F}$ and $g = d\mathbf{G}/d\mathbf{F}$, it holds

$$\inf_{G \in M_\theta(\mathbf{F})} \int_{\mathbb{R}} \varphi\left(\frac{d\mathbf{G}}{d\mathbf{F}}(x)\right) \mathbf{F}(dx) = \sup_{\xi \in \mathbb{R}^l} - \int_{\mathbb{R}} \psi\left(\langle \xi, g(x,\theta)\rangle\right) \mathbf{F}(dx).$$

Secondly, if $C = g(F^{-1}(.),\theta)$, $a = 0$, $\mu = \lambda$ and $g = dT/d\lambda$, it holds

$$\inf_{T \in M'_\theta(\mathbf{F})} \int_0^1 \varphi\left(\frac{dT}{d\lambda}\right) d\lambda = \sup_{\xi \in \mathbb{R}^l} - \int_0^1 \psi\left(\langle \xi, g(F^{-1}(u),\theta)\rangle\right) \lambda(du).$$

Lemma 1.2 concludes the proof. □

The estimators of minimum divergence used in [82] and [29] can be expressed in terms of T , introducing the empirical distribution of the sample in place of the true unknown distribution $\mathbf{F}_{\theta_0}$. For each θ in Θ it holds

$$\inf_{\mathbf{G} \in M_\theta(\mathbf{F}_n)} \int_{\mathbb{R}} \varphi\left(\frac{d\mathbf{G}}{d\mathbf{F}_n}\right) \mathbf{F}_n(dx) = \inf_{T \in M'_\theta(\mathbf{F}_n)} E_1(T)$$

and

$$\theta_n := \arg \inf_{\theta\in\Theta} \inf_{T\in M'_\theta(\mathbf{F}_n)} E_1(T).$$

Remark 1.13. *Note that if* $T \in M'_\theta(\mathbf{F}_n)$, $T : [0;1] \to [0;1]$ *is* λ*-a.e. derivable and verifies*

$$\sum_{i=1}^{n-1} g(x_{i:n}, \theta) \left(T\left(\frac{i+1}{n}\right) - T\left(\frac{i}{n}\right)\right) = 0.$$

The plug-in estimator that realizes the minimum of the divergence between a given distribution and the model constrained by moment equations results from the minimum of an energy of deformation when the deformation is constrained on the weights of the empirical distribution but is defined on almost all the cumulated weights on $[0;1]$.
We will now see that the approach of Section 1.4.2 consists in minimizing a deformation of the points of the distribution of interest instead of the weights.

The case of models defined by L-moment constraints

Similarly as for the case of models defined by moment constraints we now see that the solution of the minimum divergence problem (primal problem) holds without assuming $\mathbf{F}^{-1}$ absolutely continuous with respect to the Lebesgue measure.

Proposition 1.6. *Let* F *and* G *be two arbitrary cdf's. Let* $L''_\theta(\mathbf{F}^{-1})$ *denote the class of all functions* T *which are a.e derivable on* $\mathbb{R}$ *defined through*

$$L''_\theta(\mathbf{F}^{-1}) = \left\{ T : \mathbb{R} \to \mathbb{R} \text{ s.t. } a_\varphi < \frac{dT}{d\lambda} < b_\varphi \ \lambda - a.e. \text{ and } \int_{\mathbb{R}} K(F(x)) \frac{dT}{d\lambda}(x)\lambda(dx) = f(\theta) \right\}. \tag{1.30}$$

Then, with $L_\theta(\mathbf{F}^{-1})$ *defined in (1.25)*

$$\inf_{G^{-1}\in L_\theta(\mathbf{F}^{-1})} \int_0^1 \varphi\left(\frac{d\mathbf{G}^{-1}}{d\mathbf{F}^{-1}}(u)\right) \mathbf{F}^{-1}(du) = \inf_{T\in L''_\theta(\mathbf{F}^{-1})} \int_{\mathbb{R}} \varphi\left(\frac{dT}{d\lambda}\right) d\lambda.$$

Proof. This results from a combination of Corollaries 1.1 and 1.2. (1.30) is well defined. □

In the following, we consider the estimator of θ

$$\hat{\theta}_n = \arg \inf_{\theta\in\Theta} \inf_{T\in L''_\theta(\mathbf{F}_n^{-1})} \int_{\mathbb{R}} \varphi\left(\frac{dT}{d\lambda}\right) d\lambda. \tag{1.31}$$

The estimator $\hat{\theta}_n$ defined in (1.31) coincides with (1.22) thanks to the above Proposition 1.6.

Remark 1.14. $\cup_\theta L_\theta(\mathbf{F}^{-1})$ *and* $\cup_\theta L''_\theta(\mathbf{F}^{-1})$ *both represent the same model with L-moments constraints, seen through a reference measure* $\mathbf{F}^{-1}$*. This model is either expressed as the space of quantile measures absolutely continuous with respect to* $\mathbf{F}^{-1}$ *satisfying the L-moment constraints or as the space of all deformations* $\mathbf{F}^{-1} \to T \circ F^{-1}$ *of the reference measure* $\mathbf{F}^{-1}$ *such that the deformed measure satisfies the L-moment constraints. In the second point of view* T *is derivable* λ*-a.e. even if the reference measure is* $\mathbf{F}_n^{-1}$*.*

Remark 1.15. *For the set of deformations $L''_\theta(\mathbf{F}_n^{-1})$ (whenever non void), the duality for finite distributions is expressed through the following equality :*

$$\inf_{T\in L''_\theta(\mathbf{F}_n^{-1})} \int \varphi\left(\frac{dT}{d\lambda}\right) d\lambda = \sup_{\xi\in\mathbb{R}^l} \xi^T f(\theta) - \sum_{i=1}^{n-1} \psi\left(\xi^T K\left(\frac{i}{n}\right)\right)(x_{i+1:n} - x_{i:n}).$$

Remark that we incorporate the property of any T_1 in the model $L''_\theta(\mathbf{F}_n^{-1})$ to verify $a_\varphi < \frac{dT_1}{d\lambda} < b_\varphi$ λ-a.s.

Example 1.12. *If we take the χ^2-divergence $\varphi(x) = \frac{(x-1)^2}{2}$, then $\psi(t) = \frac{1}{2}t^2 + t$ and the solution ξ_1^* of the equation (1.28) is*

$$\xi_1^* = \Omega^{-1}\left(f(\theta) - \int K(F(x))d\lambda\right)$$

with

$$\Omega = \int K(F(x))K(F(x))^T d\lambda.$$

If we set $\Omega_n = \int K(F_n(x))K(F_n(x))d\lambda$, the estimator shares similarities with the GMM estimator

$$\hat\theta_n = \arg\inf_{\theta\in\Theta}\left(f(\theta) - \int K(F_n(x))d\lambda\right)\Omega_n^{-1}\left(f(\theta) - \int K(F_n(x))d\lambda\right).$$

This divergence should thus be favored for its fast implementation.

Remark 1.16. *We did not consider the constraints of positivity classically assumed in moment estimating equations for the sake of simplicity of dual representations. We could suppose that the transformation T is an increasing mapping. It would be the case if, for example, the divergence chosen is the Kullback-Leibler one. Indeed, in this case, problem (1.31) is well defined since $\varphi(x) = +\infty$ for all $x \le 0$.*

1.6.2 Transportation functionals and multivariate generalization

The notion of a deformation which was introduced in the above section is close to the notion of a transportation. The reformulation presented in Proposition 1.6 calls for a natural extension in this respect. Let us recall the definition of a transportation in $\mathbb{R}$.

Definition 1.4. *The pushforward measure of $\mathbf{F}$ through T is the measure denoted by $T\#\mathbf{F}$ satisfying*

$$T\#\mathbf{F}(B) = \mathbf{F}(T^{-1}(B)) \text{ for every Borel subset } B \text{ of } \mathbb{R}$$

T is said to be a transportation map between $\mathbf{F}$ and $\mathbf{G}$ if $T\#\mathbf{F} = \mathbf{G}$. If X and Y are associated with respective cdf F and G then $T(X) =_d Y$.

We write L_θ (equation (1.17)) as a space of positive measures

$$L_\theta = \left\{\mathbf{G} \in M \text{ s.t. } \int_0^1 L(u)G^{-1}(u)du = -f(\theta)\right\}.$$

It follows that an alternative to the estimator (1.31) may be defined by

$$\hat\theta_n^{(tr)} = \arg\inf_{\theta\in\Theta} \inf_{T:T\#\mathbf{F}_n\in L_\theta} E_1(T) \tag{1.32}$$

where $\mathbf{F}_n$ is the empirical measure on the observed sample $x_1, ..., x_n$ and where $E_1(T) := \int_{\mathbb{R}} \varphi\left(\frac{dT}{d\lambda}\right) d\lambda$ stands for the energy which transports $\mathbf{F}_n$ onto some $\mathbf{G}$.

Remark 1.17. *The study of estimators (1.32) is beyond the scope of this work. The advantage of $\hat{\theta}_n^{(tr)}$ over $\hat{\theta}_n$ given by equation (1.32) is the absence of absolute continuity assumption in the model L_θ. The estimation process may gain in generality.*

In transportation theory, it is customary to define a cost function instead of an energy function. Given a convex cost function $c : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, a second alternative version to (1.31) is

$$\hat{\theta}_n = \arg \inf_{\theta \in \Theta} \inf_{T : T\#\mathbf{F}_n \in L_\theta} \int_{\mathbb{R}} c(x, T(x)) \mathbf{F}_n(dx). \tag{1.33}$$

Remark 1.18. *Whereas the estimator given by Equation (1.31) minimizes an energy expressed in function of T' (the estimation process then penalizes big values of T'), the optimal transportation estimations depends on the function T itself and penalizes the distance between each x_i and $T(x_i)$ i.e. the "initial" state and the deformed state.*

Example 1.13. *Let us present an example of an estimator stemming from the optimal transportation problem (1.33) in the context of models constrained by L-moments equations.*
Consider the cost function $c(x, y) = (x - y)^2$. Then the transportation problem reduces to

$$\inf_{T : T\#dF_n \in L_\theta} \int_{\mathbb{R}} (x - T(x))^2 \mathbf{F}_n(dx) = \inf_{\mathbf{G} \in L_\theta} W_2(\mathbf{F}_n, \mathbf{G})^2 := \inf_{\mathbf{G} \in L_\theta} \int_0^1 |F_n^{-1}(t) - G^{-1}(t)|^2 dt,$$

and W_2 is called the Wasserstein distance (see e.g. [108]). The estimator (1.33) will then be defined by

$$\begin{aligned} \hat{\theta}_n &:= \arg \inf_{\theta \in \Theta} \min_{\mathbf{G} \in L_\theta} W_2(\mathbf{F}_n, \mathbf{G}) \\ &= \arg \inf_{\theta \in \Theta} \min_{y \in E'_n} \frac{1}{n} \sum_{i=1}^{n} |x_{i:n} - y_i|^2 \\ &= \arg \inf_{\theta \in \Theta} \sum_{r=2}^{l} (f_r(\theta) - l_r)^2 \end{aligned}$$

with l_r given by equation (1.14).

This estimation results as a quadratic projection of the l first L-moments of the empirical distribution onto the space formed by the constraints induced by the model.

As transportation is well defined for measure in $\mathbb{R}^d$ in contrast with quantile measures, this may appear as a way to generalize L-moments constrained models and associated estimators of the form (1.32) ; we could also consider estimators of the form (1.33), importing henceforth optimal transportation concepts in the field of multivariate quantile models ; see Chapter 2.

1.6.3 Relation to elasticity theory

It may be of interest for the statistician to observe that, besides the probabilistic context of semiparametrics, the minimization of a φ divergence over a class of functions defined by L-moments (see (1.31)) is in the same vein as finding the deformation of a solid under a given force L and given boundary constraints. Let us consider a solid defining a domain $\Omega \subset \mathbb{R}^3$. This solid can be deformed under the action of volumetric or surface forces. This deformation can be described by a function $T : \Omega \to \mathbb{R}^3$. The deformed solid

will be defined on the volume $T(\Omega)$. The gradient of deformation is then ∇T.
The general equations describing the equilibrium of the solid under volumetric forces L defined on Ω read (we omit boundary forces)

$$-\text{div}S = L$$

where S is a tensor describing the configuration of the solid [24]. Hyper-elasticity is often assumed i.e. the solid is supposed to dissipate no energy during the deformation. In mathematical terms, this means the existence of a function W such that

$$S(T) = \frac{\partial W}{\partial T}(T).$$

From these above relations, the energy of deformation is expressed on the form [74][24]

$$\mathcal{E}(T) = \int_\Omega W(\nabla T(x))dx - \int_\Omega L(x).T(x)dx.$$

W is usually convex and represents physical properties of the solid. It is then customary in mechanical physics to assume the principle of least action and to study the T minimizing the variational problem

$$\inf_{T \text{ admissible}} \mathcal{E}(T).$$

The space of admissible T describes the constraints, such as boundary conditions. If we could write the volumetric force term (namely the right hand side of $\mathcal{E}(T)$) as fixed constraints, we remark similarities with the estimation given by equation 1.31

$$\hat{\theta}_n = \arg \inf_{\theta \in \Theta} \inf_{\int_\Omega L(x).T(x)dx = f(\theta)} \int_\Omega \varphi\left(\nabla T(x)\right) dx.$$

Moreover, microscopic and macroscopic scales can be related through convergence results. Let us present the microscopic models of the same solid represented by N particles $x_1, ..., x_N$, corresponding for example to the intersection of Ω with a lattice of scale ϵ. If V denotes an interaction potential, the energy of the solid subjected to a deformation T would be

$$\mathcal{E}_N(T) = \frac{\epsilon^3}{2} \sum_{i=1}^{N} \sum_{j \neq i} V\left(\frac{T(x_i) - T(x_j)}{\epsilon}\right) - \sum_{i=1}^{N} L(x_i).T(x_i)$$

Under some assumptions (see [23]), it can be proved that if $\epsilon \to 0$ (i.e. $N \to \infty$), then

$$\mathcal{E}_N(T) \to_{N \to \infty} \mathcal{E}(T).$$

This short account may give us some intuition about the present estimation procedure.

1.7 Asymptotic properties of the L-moment estimators

In this section, we study the convergence of the estimator given by the equation (1.31). The proof of the two asymptotic theorems are postponed to the Appendix.

Theorem 1.2. *Let $x_1, ..., x_n$ be an observed sample drawn iid from a distribution F_0 with finite variance. Let us suppose that*

- *there exists θ_0 such that $F_0 \in L_{\theta_0}$, θ_0 is the unique solution of the equation $f(\theta) = f(\theta_0)$*

- *f is continuous and $\Theta \subset \mathbb{R}^d$ is compact*
- *the matrix $\Omega_0 = \int K(F_0(x))K(F_0(x))^T dx$ is non singular.*

Then,

$$\hat{\theta}_n \to \theta_0 \text{ in probability as } n \to \infty.$$

We may now turn to the limit distribution of the estimator. Let

- $J_0 = J_f(\theta_0)$ be the Jacobian of f with respect to θ in θ_0
- $M = (J_0^T \Omega^{-1} J_0)^{-1}$
- $H = M J_0^T \Omega^{-1}$
- $P = \Omega^{-1} - \Omega^{-1} J_0 M J_0^T \Omega^{-1}$

Theorem 1.3. *Let $x_1, ..., x_n$ be an observed sample drawn iid from a distribution F_0 with finite variance. We assume that the hypotheses of Theorem 1.2 holds. Moreover, we assume that*

- *$\theta_0 \in int(\Theta)$*
- *J_0 has full rank*
- *f is continuously differentiable in a neighborhood of θ_0*

Then,

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_n - \theta_0 \\ \hat{\xi}_n \end{pmatrix} \to_d \mathcal{N}_{d+l} \left(0, \begin{pmatrix} H\Sigma H^T & 0 \\ 0 & P\Sigma P^T \end{pmatrix} \right)$$

The estimator of the minimum of the divergence from $\mathbf{F}$ onto the model, namely $2n \left[\hat{\xi}_n^T f(\hat{\theta}_n) - \int \psi(\hat{\xi}_n^T K(F_n(x))dx \right]$, does not converge to a χ^2-distribution as in the case of moment condition models [82]. However, we can state an alternative result.

Corollary 1.3. *Let us assume that the hypotheses of Theorem 1.3 hold.*
Let us consider $S_n = n\hat{\xi}_n^T (P_n \Sigma_n P_n^T)^{-1} \hat{\xi}_n$ with P_n and Σ_n the respective empirical versions of P and Σ.
If $P\Sigma P$ is non singular then

$$S_n \to_d \chi^2(l)$$

where $\chi^2(l)$ denotes a chi-square distribution with l degrees of freedom.

Proof. From Theorem 1.3, we have that

$$n^{1/2} \hat{\xi}_n \to_d X = \mathcal{N}_l(0, P\Sigma P)$$

where X denotes such a multivariate Gaussian random vector.
Furthermore

$$P_n \Sigma_n P_n \to_p P\Sigma P.$$

Hence, for n large enough, $P_n \Sigma_n P_n$ is invertible and by Slutsky Theorem

$$n\hat{\xi}_n^T (P_n \Sigma_n P_n)^{-1} \hat{\xi}_n \to_p X^T X =_d \chi^2(l).$$

□

Since the weak convergence of S_n to a chi-square is independent of the value of θ_0, this result may be used in order to build confidence regions related to the semi-parametric model.

1.8 Numerical applications : Inference for Generalized Pareto family

1.8.1 Presentation

The Generalized Pareto Distributions (GPD) are known to be heavy-tailed distributions. They are classically parametrized by a location parameter m, which we assume to be 0, a scale parameter σ and a shape parameter ν. They can be defined through their density :

$$f_{\sigma,\nu}(x) = \begin{cases} \frac{1}{\sigma}\left(1+\nu\frac{x}{\sigma}\right)^{-1-1/\nu}\mathbb{1}_{x>0} & \text{if } \nu > 0 \\ \frac{1}{\sigma}\exp\left(\frac{x}{\sigma}\right)\mathbb{1}_{x>0} & \text{if } \nu = 0 \\ \frac{1}{\sigma}\left(1+\nu\frac{x}{\sigma}\right)^{-1-1/\nu}\mathbb{1}_{-\sigma/\nu>x>0} & \text{if } \nu < 0 \end{cases}$$

Let us remark that if $\nu \geq 1$, the GPD does not have a finite expectation. We perform different estimations of the scale and the shape parameter of a GPD from samples with size $n = 100$.
We will estimate the parameters in the model composed by the distributions of all r.v's X whose second, third and fourth L-moments verify

$$\begin{cases} \lambda_2 &= \frac{\sigma}{(1-\nu)(2-\nu)} \\ \frac{\lambda_3}{\lambda_2} &= \frac{1+\nu}{3-\nu} \\ \frac{\lambda_4}{\lambda_2} &= \frac{(1+\nu)(2+\nu)}{(3-\nu)(4-\nu)} \end{cases} \tag{1.34}$$

for any $\sigma > 0, \nu \in \mathbb{R}$. These distributions share their first L-moments with those of a GPD with scale and shape parameter σ and ν (see [64]). This estimation will be compared with classical parametric estimators detailed hereafter.

1.8.2 Moments and L-moments calculus

The variance and the skewness of the GPD are given by

$$\begin{cases} var &= \mathbb{E}[(X-\mathbb{E}[X])^2] = \frac{\sigma^2}{(1-\nu)^2(1-2\nu)} \\ t_3 &= \mathbb{E}[\left(\frac{X-\mathbb{E}[X]}{\mathbb{E}[(X-\mathbb{E}[X])^2]}\right)^3] = \frac{2(1+\nu)\sqrt{1-2\nu}}{1-3\nu} \end{cases}$$

Let us remark that var and t_3 respectively exist since $\nu < 1/2$ and $\nu < 1/3$.
On the other hand, the first L-moments are given by equation 1.34. Assuming $\nu < 1$ entails existence of the L-moments.

1.8.3 Simulations

We perform $N = 500$ runs of the following estimators

- the proposed estimation (equation (1.31)) for the χ^2-divergence and the modified Kullback (KL_m) divergence with the constraints estimated on the L-moments of order $2, 3, 4$
- the estimate defined through the L-moment method, based on the empirical second L-moment $\hat{\lambda}_2$ and the fourth L-moment ratio $\hat{\tau}_4 = \frac{\lambda_4}{\lambda_2}$

$$\hat{\nu} = \frac{7\hat{\tau}_4 + 3 - \sqrt{(\hat{\tau_4}^2 + 98\hat{\tau}_4 + 1}}{2(\hat{\tau}_4 - 1)}$$
$$\hat{\sigma} = \hat{\lambda}_2(1-\hat{\nu})(2-\hat{\nu})$$

		$n = 30$			$n = 100$		
Estimation method	Parameter	Mean	Median	StD	Mean	Median	StD
χ^2-divergence	σ	4.68	4.41	2.52	3.80	3.75	0.90
KL_m-divergence	σ	6.44	4.77	8.02	4.08	3.95	4.00
L-moment method	σ	5.67	4.98	3.44	3.96	3.80	1.09
Moment method	σ	17.17	10.45	62.95	17.15	11.64	19.52
MLE	σ	3.33	3.17	1.14	3.08	3.07	0.57
χ^2-divergence	ν	0.38	0.39	0.24	0.55	0.55	0.16
KL_m-divergence	ν	0.37	0.38	0.24	0.38	0.37	0.16
L-moment method	ν	0.33	0.38	0.31	0.54	0.56	0.18
Moment method	ν	0.08	0.12	0.12	0.21	0.22	0.06
MLE	ν	0.61	0.63	0.33	0.68	0.69	0.17

TABLE 1.1 – Estimates of GPD scale and shape parameters for $\nu = 0.7$ and $\sigma = 3$ (the moment method has little sense since $\nu > 0.5$) for the first scenario without outliers

- the estimate defined through the moment method estimated from the empirical variance $\hat{var}$ and skewness $\hat{t}_3$

$$\hat{\nu} = \frac{2(1+\hat{t}_3)\sqrt{1-2\hat{t}_3}}{1-3\hat{t}_3}$$
$$\hat{\sigma} = \sqrt{\hat{var}(1-\hat{t}_3)^2(1-2\hat{t}_3)}$$

- the MLE defined in the GPD family

We present the following different features for any of the above estimators

- the mean of the N estimates based on the N runs
- the median of the N estimates based on the N runs
- the standard deviation of the N estimates
- the L_1 distance between the estimated generalized Pareto density and the true density, namely

$$\int_{x\geq 0} |f_{\hat{\sigma},\hat{\nu}}(x) - f_{\sigma,\nu}(x)|dx$$

which, by Scheffé Lemma, equals twice the maximum error committed substituting $f_{\sigma,\nu}$ by $f_{\hat{\sigma},\hat{\nu}}$

$$\int_{x\geq 0} |f_{\hat{\sigma},\hat{\nu}}(x) - f_{\sigma,\nu}(x)|dx = 2 \sup_{A\in\mathcal{B}(\mathbb{R})} \left| \int_A f_{\hat{\sigma},\hat{\nu}}(x) - \int_A f_{\sigma,\nu}(x)|dx \right|.$$

Finally, we present four different scenarios which illustrate robusness properties of any of the above estimators, as well as their behavior under misspecification :

- a first scenario without outliers : samples of size 30 or 100 are drawn from a GPD
- two more scenarios with 10% outliers : samples of size 27 or 90 are drawn from a GPD. The remaining points are drawn from a Dirac the value of which depends on the shape parameter
- a fourth scenario without outliers but with misspecification : samples of size 30 or 100 are drawn from a Weibull distribution.

Unsurprisingly, the MLE performs well under the model and the L-moment method has an overall better behavior than the classical moment method for the considered heavy-tailed distributions (see Table 1.1). Furthermore, we observe that the χ^2- divergence is

		$n = 30$			$n = 100$		
Estimation method	Parameter	Mean	Median	StD	Mean	Median	StD
χ^2-divergence	σ	12.43	12.24	2.83	12.29	12.21	1.62
KL_m-divergence	σ	24.01	19.36	49.38	27.30	20.99	48.75
L-moment method	σ	22.27	20.83	5.69	21.68	21.03	3.09
Moment method	σ	80.97	76.27	20.89	80.93	76.84	31.09
MLE	σ	3.06	2.88	1.08	2.88	2.86	0.55
χ^2-divergence	ν	0.55	0.55	0.05	0.54	0.54	0.04
KL_m-divergence	ν	0.50	0.52	0.24	0.54	0.49	0.27
L-moment method	ν	0.54	0.54	0.06	0.54	0.53	0.04
Moment method	ν	0.07	0.08	0.02	0.08	0.07	0.03
MLE	ν	1.48	1.44	0.22	1.50	1.49	0.11

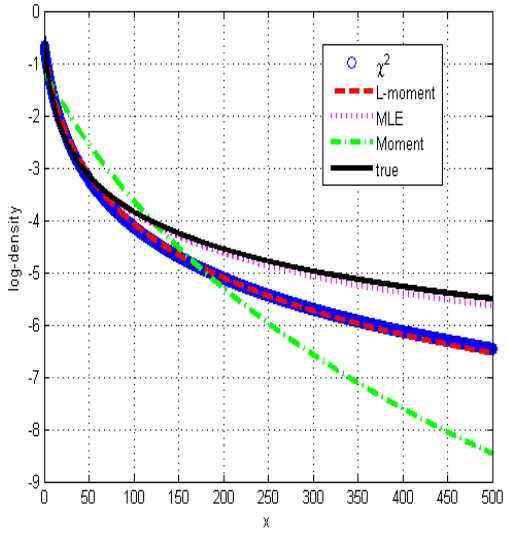
Table 1.2 – Estimates of GPD scale and shape parameters for $\nu = 0.7$ and $\sigma = 3$ for a sample with 10% outliers of value 300 (the moment method has little meaning since $\nu > 0.5$)

		$n = 30$			$n = 100$		
Estimation method	Parameter	Mean	Median	StD	Mean	Median	StD
χ^2-divergence	σ	4.32	4.23	0.91	4.45	4.42	0.51
KL_m-divergence	σ	5.04	4.90	1.15	5.07	5.08	0.67
L-moment method	σ	5.18	5.04	1.44	5.11	5.04	0.75
Moment method	σ	8.64	8.44	0.92	8.54	8.48	0.50
MLE	σ	3.12	3.08	0.87	3.08	3.05	0.49
χ^2-divergence	ν	0.27	0.28	0.08	0.27	0.27	0.05
KL_m-divergence	ν	0.25	0.25	0.09	0.24	0.24	0.05
L-moment method	ν	0.24	0.24	0.10	0.24	0.24	0.06
Moment method	ν	0.01	0.02	0.04	0.01	0.02	0.02
MLE	ν	0.56	0.54	0.17	0.55	0.55	0.09

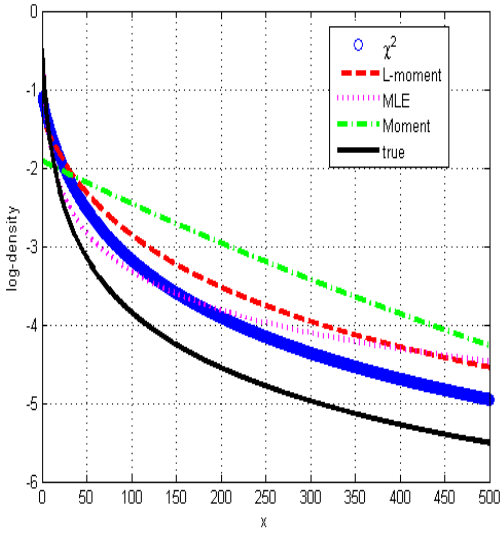
Table 1.3 – Estimates of GPD scale and shape parameters for $\nu = 0.1$ and $\sigma = 3$ for a sample with 10% outliers of value 30

	$n = 30$				$n = 100$			
Estimation method	Sc 1	Sc 2	Sc 3	Sc 4	Sc 1	Sc 2	Sc 3	Sc 4
χ^2-divergence	2.53	7.20	3.16	2.63	1.55	7.32	3.28	1.80
L-moment method	3.10	10.07	4.09	4.31	1.70	9.93	4.07	3.51
Moment method	6.79	14.47	7.07	8.69	6.91	14.42	6.98	9.98
MLE	1.78	2.83	2.68	11.69	0.97	2.42	2.33	9.25

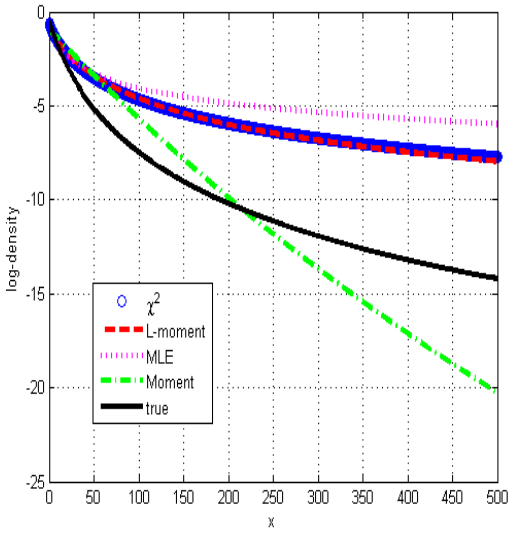
Table 1.4 – L_1-distances (to be multiplied by 10^{-4}) between GPD densities for different scenarios; Scenario (Sc) 1 corresponds to a simulated GPD with $\nu = 0.7$ and $\sigma = 3$; Scenario 2 corresponds to a simulated GPD with $\nu = 0.7$, $\sigma = 3$ and 10% outliers of value 300; Scenario 3 corresponds to a simulated GPD with $\nu = 0.1$, $\sigma = 3$ and 10% outliers of value 30; Scenario 4 corresponds to a simulated Weibull distribution with $\nu = 0.4$ and $\sigma = 3$



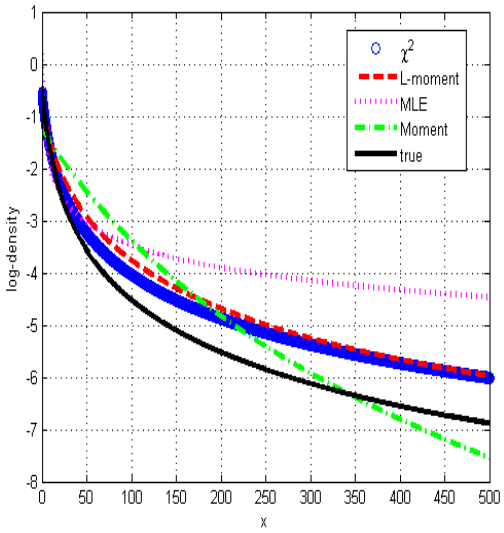
(a) Simulated GPD with $\nu = 0.7$ and $\sigma = 3$



(b) Simulated GPD with $\nu = 0.7$, $\sigma = 3$ and 10% outliers of value 300



(c) Simulated GPD with $\nu = 0.1$, $\sigma = 3$ and 10% outliers of value 30



(d) Simulated Weibull distribution with $\nu = 0.4$ and $\sigma = 3$

FIGURE 1.2 – Estimated GPD densities with estimated parameters for simulated scenarios (with a logarithmic scale)

more robust than the modified Kullback as indeed expected.
The interesting result lies in their behavior with outliers and misspecification. Indeed, we can see that L-moment-based estimators perform well on the shape parameter whereas the MLE provides a good estimation of the scale parameter but overestimates the shape parameter. In that sense, the L-moments method can be used for the robust estimation of the shape parameter of a GPD in case of contamination by outliers. However, even with outliers, the MLE performs well in term of L_1-distance computed on the estimated densities. It is under misspecification that the performance of the MLE drops as measured

by the L_1 criterion. This confirms the flexibility of models defined only through moment or L-moment equations that are less dependent on the GPD model.
Moreover, the L_1-distance between the model and its estimation has an order between 10^{-3} and 10^{-4}. The error committed by the estimation under models defined through L-moments conditions is the most stable over the proposed scenarios. We can then affirm that we can estimate the probability of events if the true value of this probability is of order 10^{-3} (the error of estimation for the estimator based on L-moments method would approximately be of 30% depending on the size of the sample and the scenario).

1.9 Appendix

1.9.1 Proof of Lemma 1.1

Let $x \in \mathbb{R}$. We denote by F the cdf of X and by A_t the event

$$A_t = \{x \in \mathbb{R} \text{ s.t. } F(x) \geq t\}$$

We then have $Q(t) = \inf A_t$. We wish to prove :

$$\{t \in [0;1] \text{ s.t. } Q(t) \leq x\} = \{t \in [0;1] \text{ s.t. } t \leq F(x)\} \tag{1.35}$$

We temporarily admit this assertion. Then

$$\mathbb{P}[Q(U) \leq x] = \mathbb{P}[U \leq F(x)] = F(x)$$

which ends the proof. It remains to prove (2.17).
First, the definition of Q yields

$$\{t \leq F(x)\} \Rightarrow \{x \in A_t\} \Rightarrow \{Q(t) \leq x\}.$$

Secondly, let t be such that $Q(t) \leq x$. Then by monotonicity of F, $F(Q(t)) \leq F(x)$. We then claim that

$$Q(t) \in A_t.$$

Indeed, let us suppose the contrary and consider a strictly decreasing sequence $x_n \in A_t$ such that

$$\lim_{n\to\infty} x_n = \inf A_t = Q(t).$$

By right continuity of F

$$\lim_{n\to\infty} F(x_n) = F(Q(t))$$

and, on the other hand, by definition of A_t,

$$\lim_{n\to\infty} F(x_n) \geq t$$

i.e. $Q(t) \in A_t$ which contradicts the hypothesis. Then $Q(t) \in A_t$ i.e. $t \leq F(Q(t))$ thus $t \leq F(x)$. We have proved that

$$\{Q(t) \leq x\} \Rightarrow \{t \leq F(x)\}.$$

1.9.2 Proof of Lemma 1.2

Let us recall that the support of a measure μ defined on $X \subset \mathbb{R}$ is the largest closed set $C \subset X$ such that

$$U \in B(X) \text{ and } U \cap C \neq \emptyset \Rightarrow \mu(U \cap C) > 0$$

where $B(X)$ denotes the Borel sets in X. Let S be the support of $\mathbf{F}^{-1}$. Then $[0;1]\backslash S$ is an open set in $[0;1]$ i.e. a countable union of intervals $\cup_{i\geq 1}]t_{2i}, t_{2i+1}[$ and

$$\begin{aligned}
\int_0^1 a(F^{-1}(t))dt &= \int_S a(F^{-1}(t))dt + \sum_{i\geq 1}\int_{]t_{2i};t_{2i+1}[} a(F^{-1}(t))dt \\
&= \int_{F^{-1}(S)} a(x)dF(x) + \sum_{i\geq 1} a(F^{-1}(t_{2i}))(t_{2i+1}-t_{2i}) \\
&= \int_{F^{-1}(S)} a(x)dF(x) + \sum_{i\geq 1}\int_{\{F^{-1}(t_{2i})\}} a(x)dF(x) \\
&= \int_{F^{-1}(S)\cup\left(\cup_{i\geq 1}\{F^{-1}(t_{2i})\}\right)} a(x)dF(x).
\end{aligned}$$

The second equality stems from the definition of the quantile as left-continuous function and from the fact that F^{-1} is strictly monotone on S.
As F^{-1} is constant on the open interval $[t_{2i};t_{2i+1}[$, $\{F^{-1}(t_{2i})\} = F^{-1}([t_{2i};t_{2i+1}[)$. Hence

$$\begin{aligned}
F^{-1}(S) \cup \left(\cup_{i\geq 1}\{F^{-1}(t_{2i})\}\right) &= F^{-1}([0;1]) \\
&= \{x \in \mathbb{R} \text{ s.t. there exists } t \text{ with } F^{-1}(t) = x\} = supp(F).
\end{aligned}$$

We conclude the first part of the proof since

$$\int_{supp(F)} a(x)dF(x) = \int_{\mathbb{R}} a(x)dF(x).$$

The second part of the proof can be proved similarly since the above arguments are not particular to a specific measure.

1.9.3 Proof of Proposition 1.4

The proof is directly adapted from the proof of Theorem II.2 of Csiszár et al. [43].
Let us begin with the fundamental lemma inspired from Theorem 2.9 of Borwein and Lewis[25].

Lemma 1.3. *Let $C : \Omega \to \mathbb{R}^l$ be an array of bounded functions such that*

$$\int_\Omega \|C(x)\|d\mu(x) < \infty.$$

We denote

$$L_{C,a} = \left\{g \ s.t. \ \int_\Omega g(t)C(t)d\mu(t) = a\right\}.$$

If there exists some g in $L_{C,a}$ such that $a_\varphi < g < b_\varphi$ μ-a.s and $\int_\Omega \|g(t)C(t)\|d\mu(t) < \infty$, then there exists $a'_\varphi > a_\varphi$, $b'_\varphi < b_\varphi$ and $g_b \in L_{C,a}$ such that $a'_\varphi \leq g_b(x) \leq b'_\varphi$ for all $x \in \Omega$.

Proof. Let L denotes the subspace of $\mathbb{R}^l$ composed by the vectors representable as $\int_\Omega g C d\mu$ for some $g : \Omega \to \mathbb{R}^l$. Let us denote by a_n a decreasing sequence $a_n \to a_\varphi$, by b_n a increasing one $b_n \to b_\varphi$ and let T_n be the set

$$T_n = \{x \in \Omega \text{ s.t. } a_n \leq g(x) \leq b_n\}.$$

We first claim that, for n large enough

$$L = L_n = \left\{ \int_\Omega h C d\mu \text{ with } h(x) = 0 \text{ if } x \notin T_n \text{ and } h \text{ bounded} \right\}.$$

Indeed, if not, we can build a sequence of vectors v_n such that $\|v_n\| = 1$, $v_n \in L^\perp$ and $v_n \to v \in L$. Furthermore, $v_n \in L^\perp$ means

$$\langle v_n, \int_\Omega h C d\mu \rangle = \int_\Omega h \langle v_n, C \rangle d\mu = 0$$

then $\langle v_n, C \rangle = 0$ for all $x \in T_n$ μ-a.s. Hence $\langle v, C \rangle = 0$ μ-a.s. and $v \in L^\perp$ which contradicts $v \in L$ with $\|v\| = 1$.
Let us then fix some n_0 such that $L_{n_0} = L$. We denote by

$$L_n(\delta) = \left\{ \int_\Omega h C d\mu \text{ with } h(x) = 0 \text{ if } x \notin T_n \text{ and } |h(x)| < \delta \text{ for } x \in \Omega \right\}.$$

Then, the affine hull of $L_n(\delta)$ is the vector space L and $0 \in L_n(\delta)$. We can consider the function g_n

$$g_n(x) = \begin{cases} a_n & \text{if} \quad g(x) < a_n \\ g(x) & \text{if} \quad b_n \leq g(x) \leq a_n \\ b_n & \text{if} \quad g(x) > b_n \end{cases}$$

Then $\| \int_\Omega (g_n - g) C d\mu \| \to_{n \to \infty} 0$. Indeed we can apply the dominated convergence theorem since, for any $x \in \Omega$, $g_n(x) \to g$ and

$$\begin{aligned} \|(g_n(x) - g(x))C(x)\| &= \|\mathbb{1}_{g(x)<a_n}(a_n - g(x))C(x) + \mathbb{1}_{g(x)>b_n}(g(x) - b_n)C(x)\| \\ &\leq (\|a_0 - g(x)\| + \|b_0 - g(x)\|)\|C(x)\| \\ &\leq (\|a_0\| + \|b_0\|)\|C(x)\| + 2\|g(x)\|\|C(x)\| \end{aligned}$$

which is μ-measurable by hypothesis.
We conclude that $\int_\Omega (g_n - g) C d\mu \in L_{n_0}(\delta)$ for n large enough because $0 \in L_{n_0}(\delta)$. Hence there exists h such that $\int_\Omega (g_n - g) C d\mu = \int_\Omega h C d\mu$, $|h(x)| = 0$ for $x \notin T_{n_0}$ and $|h(x)| < \delta$ for x in T_{n_0}.
Therefore for $x \in \Omega$, $\min(a_n, a_{n_0} - \delta) \leq g_n(x) + h(x) \leq \min(b_n, b_{n_0} + \delta)$ and $\int_\Omega (g_n + h) C d\mu = \int_\Omega g C d\mu$. As δ is arbitrarily small, h is the null function. □

We can now prove the duality equality. Let note for $c \in \mathbb{R}^l$ $I(c) = \inf_{\int g C d\mu = c} \int_\Omega \varphi(g) d\mu$ and

$$J(c) = \begin{cases} 0 & \text{if } c = a \\ +\infty & \text{otherwise} \end{cases}.$$

Then

$$\inf_{g \in L_{C,a}} \int \varphi(g) d\mu = \inf_{c \in \mathbb{R}^l} I(c) + J(c).$$

Recall that the Fenchel duality theorem ([91] p327) states that if $ri(dom(I)) \cap ri(dom(J)) \neq \emptyset$ then

$$\inf_{c \in \mathbb{R}^l} I(c) + J(c) = \max_{\xi \in \mathbb{R}^l} -I^*(\xi) - J^*(-\xi).$$

We prove that $ri(dom(I)) \cap ri(dom(J)) \neq \emptyset$. Note that $ri(dom(J)) = \{a\}$. It suffices then to prove that a belongs to $int(dom(I))$ for the topology induced by L. By the above Lemma 1.3 there exists g_b such that $a_\varphi < a'_\varphi \leq g_b(x) \leq b'_\varphi < b_\varphi$ for all $x \in \Omega$. Since $a + L_n(\delta)$ is a neighborhood of a included in $dom(I)$ for δ sufficiently small, it holds that $a \in int(L_n(\delta)) \subset int(dom(I))$.
It remains now to compute the conjugates of I and J .

$$\begin{aligned} I^*(\xi) &= \sup_{c\in\mathbb{R}^l} \langle \xi, c\rangle - \inf_{g, \int gCd\mu = c} \varphi(g)d\mu \\ &= \sup_{c\in\mathbb{R}^l} \sup_{g, \int gCd\mu = c} \langle \xi, c\rangle - \varphi(g)d\mu \\ &= \sup_g \langle \xi, \int gCd\mu\rangle - \varphi(g)d\mu \\ &= \sup_g \int \langle \xi, C\rangle g - \varphi(g)d\mu \\ &= \int \psi(\langle \xi, C\rangle)d\mu \end{aligned}$$

This equality is referred to as the integral representation of I^*. The last equality can be rigorously justified (see for example [44]).
Furthermore, $J^*(-\xi) = -\langle \xi, a\rangle$ which closes the first part of the proof, namely

$$\inf_{g\in L_{C,a}} \int_\Omega \varphi(g)\, d\mu = \sup_{\xi\in\mathbb{R}^l} \langle \xi, a\rangle - \int_\Omega \psi(\langle \xi, C(x)\rangle)d\mu.$$

As we assume ψ differentiable, then $\xi \mapsto \langle \xi, a\rangle - \int_\Omega \psi(\langle \xi, C(x)\rangle)d\mu$ is differentiable as well. It follows that any critical point is the solution of

$$\int_\Omega \psi'(\langle \xi, C(x)\rangle)C(x)d\mu = a.$$

Furthermore, as φ is strictly convex, ψ is strictly concave and for $\xi, \xi' \in \mathbb{R}^l$ and $t \in [0; 1]$ it holds

$$\begin{aligned} &\langle (1-t)\xi + t\xi', a\rangle - \int_\Omega \psi(\langle (1-t)\xi + t\xi', C(x)\rangle)d\mu \\ &= \langle (1-t)\xi + t\xi', a\rangle - \int_\Omega \psi((1-t)\langle \xi, C(x)\rangle + t\langle \xi', C(x)\rangle)d\mu \\ &< (1-t)\left[\langle \xi, a\rangle - \int_\Omega \psi(\xi, C(x)\rangle)d\mu\right] + t\left[\langle \xi', a\rangle - \int_\Omega \psi(\xi', C(x)\rangle)d\mu\right] \end{aligned}$$

i.e. the functional $\xi \to \langle \xi, a\rangle - \int_\Omega \psi(\xi, C(x)\rangle)d\mu$ is strictly convex which proves the uniqueness of ξ^*.
The continuity of $a \mapsto \xi^*(a)$ comes from the implicit function theorem. If we note $D(\xi) = \int \psi'(\langle \xi, C(x)\rangle)C(x)d\mu$ then D is continuously differentiable with a Jacobian given by

$$J_D(\xi) = \int \psi''(\langle \xi, C(x)\rangle)C(x)C(x)^T d\mu$$

which is positive definite thanks to the strict convexity of ψ.

1.9.4 Proof of Theorem 1.2

The arguments of this proof and of the following one are similar to the ones given by Newey and Smith in [82] for their Theorem 3.1 ; the essential argument is a Taylor expansion of the functionals in equation (1.27).
Let begin with a lemma adapted from Theorem 6 due to Stigler [100] :

Lemma 1.4. *Let $x_1, ..., x_n$ be an observed sample drawn iid from a distribution F with finite variance. We note F_n the empirical distribution of the sample.*
Let $A : [0; 1] \to \mathbb{R}^l$ be a continuously derivable function such that A' is bounded F^{-1}-a.e. Then

$$n^{1/2}\left(\int x dA(F_n(x)) - \int x dA(F(x))\right) \to_d N(0, \Sigma_A)$$

with

$$\Sigma_A = \iint \left[F(\min(x, y)) - F(x)F(y)\right] A'(F(x))A'(F(y))^T dxdy.$$

In the following, we will note $\frac{dT}{d\lambda}(x) = T'(x)$ for all $x \in \mathbb{R}$.

First step : maximization step

Clearly, it holds

$$\inf_{T \in \cup_\theta L''_\theta(F_n)} \int_{\mathbb{R}} \varphi(T'(x))dx \leq \inf_{T \in L''_{\theta_0}(F_n)} \int_{\mathbb{R}} \varphi(T'(x))dx. \tag{1.36}$$

By Taylor-Lagrange expansion, there exists some $D > 0$ such that for n large enough and for any t in $[1 - n^{-1/4}; 1 + n^{1/4}]$

$$\varphi(t) \leq \frac{D}{2}(t - 1)^2$$

holds.

We may then majorize the RHS in (1.36) by the solution of the quadratic case. Let

$$T'_{0,n}(x) := 1 + (f(\theta_0) - m_n)^T \Omega_n^{-1} K(F_n(x))$$

where $m_n := \int K(F_n(x))dx$ and $\Omega_n := \int_{\mathbb{R}} K(F_n(x))K(F_n(x))^T dx$. As $T'_{0,n} \in L''_\theta(F_n)$, it holds

$$\inf_{T \in L''_{\theta_0}(F_n)} \int_{\mathbb{R}} \varphi(T'(x))dx \leq \int_{\mathbb{R}} \varphi(T'_{0,n}(x))dx.$$

From Lemma 1.4, we deduce that $\Omega_n \to \Omega$ in probability. As Ω is non singular, for n large enough, Ω_n is non singular and $T'_{0,n}$ is well defined.
As $\|f(\theta_0) - m_n\| = O_P(n^{-1/2})$ from Lemma 1.4 and $\|\Omega_n^{-1}\| = O_P(1)$, for almost all $x \in \mathbb{R}$,

$$T'_{0,n}(x) = 1 + O_P(n^{-1/2})$$

and we can apply a Taylor-Lagrange maximization

$$\varphi(T'_{0,n}(x)) \leq \frac{D}{2}(f(\theta_0) - m_n)^T \Omega_n^{-1} K(F_n(x))K(F_n(x))^T \Omega_n^{-1}(f(\theta_0) - m_n).$$

By integration in the above display

$$\begin{aligned}\int_{\mathbb{R}} \varphi(T'_{0,n}(x))dx &\leq \frac{D}{2}(f(\theta_0) - m_n)\Omega_n^{-1}\left[\int_{\mathbb{R}} K(F_n(x))K(F_n(x))^T dx\right]\Omega_n^{-1}(f(\theta_0) - m_n) \\ &\leq \|f(\theta_0) - m_n\|^2 \|\Omega_n^{-1}\| = O_P(n^{-1}).\end{aligned}$$

Second step : minimization step

Since Θ is compact, and φ is strictly convex, and $\theta \mapsto \inf_{T\in L''_\theta(F_n)} \int_{\mathbb{R}} \varphi(T'(x))dx$ is continuous (see Proposition 1.4), it follows that $\hat{\theta}$ is well defined and the duality equality states

$$\begin{aligned}\inf_{T\in L''_{\hat{\theta}_n}(F_n)} \int_{\mathbb{R}} \varphi(T'(x))dx &= \sup_{\xi\in\mathbb{R}^l} \xi^T f(\hat{\theta}_n) - \int \psi(\xi^T K(F_n(x)))dx \\ &\geq \xi_n^T f(\hat{\theta}_n) - \int \psi(\xi_n^T K(F_n(x)))dx\end{aligned}$$

with

$$\xi_n = n^{-1/2} \frac{f(\hat{\theta}_n) - m_n}{\|f(\hat{\theta}_n) - m_n\|}.$$

.
Therefore

$$\xi_n^T K(F_n(x)) = O_P(n^{-1/2}) \text{ for a.e } x \in \mathbb{R}.$$

By Taylor-Lagrange expansion, there exists a constant $C > 0$ such that $|\psi(x) - x| < Cx^2$ in a neighborhood of 0. Thus, for n large enough

$$\int \psi(\xi_n^T K(F_n(x)))dx - \xi_n^T m_n < C \int \xi_n^T K(F_n(x)) K(F_n(x))^T \xi_n dx = C\xi_n^T \Omega_n \xi_n$$

and

$$\inf_{T\in L''_{\hat{\theta}_n}(F_n)} \int_{\mathbb{R}} \varphi(T'(x))dx > \xi_n^T (f(\hat{\theta}_n) - m_n) - C\xi_n^T \Omega_n \xi_n.$$

Conclusion
Combining the two inequalities, we have

$$n^{-1/2}\|f(\hat{\theta}_n) - m_n\| < C\|\Omega_n\|n^{-1} + \|f(\theta_0) - m_n\|^2 \|\Omega_n^{-1}\| = O_P(n^{-1})$$

i.e. $\|f(\hat{\theta}_n) - m_n\| = O_P(n^{-1/2})$.
By Lemma 1.4, $\|m_n - f(\theta_0)\| = O_P(n^{-1/2})$. Hence, $\|f(\hat{\theta}_n) - f(\theta_0)\| = O_P(n^{-1/2})$.
Since $f(\theta) = f(\theta_0)$ has a unique solution at θ_0, $\|f(\theta) - f(\theta_0)\|$ is bounded away from zero outside some neighborhood of θ_0. Therefore $\hat{\theta}_n$ is inside any neighborhood of θ_0 with probability approaching 1 i.e $\hat{\theta}_n \to \theta_0$ in probability.

1.9.5 Proof of Theorem 1.3

First we prove that

$$\hat{\xi}_n = \arg\max_{\xi} \xi^T f(\hat{\theta}_n) - \int \psi(\xi^T K(F_n(x))dx = O_P(n^{-1/2}).$$

Consider

$$\xi_n = \arg \max_{\xi\in\mathbb{R}^l \text{ s.t. } \|\xi\|<n^{-1/4}} \xi^T f(\hat{\theta}_n) - \int \psi(\xi^T K(F_n(x))dx,$$

where the maximum is taken on a ball of radius $n^{-1/4}$. The maximum is attained because of the concavity of the functional

$$U : \xi \mapsto \xi^T f(\hat{\theta}_n) - \int \psi(\xi^T K(F_n(x))dx.$$

For all x in a neighborhood of 0, the inequality $y - \psi(y) < -Cy^2$ for some $C > 0$ holds. For n large enough, as $\|\xi_n\| < n^{-1/4}$ we can claim (as $\psi(0) = 0$)

$$\begin{aligned} 0 &\leq \xi_n^T f(\hat{\theta}_n) - \int \psi(\xi_n^T K(F_n(x))dx \\ &\leq \xi_n^T (f(\hat{\theta}_n) - m_n) - C\xi_n^T \Omega_n \xi_n \\ &\leq \|\xi_n\|.\|f(\hat{\theta}_n) - m_n\| - C\xi_n \Omega_n \xi_n, \end{aligned}$$

with $m_n := \int K(F_n(x))dx$.
Furthermore, there exists $D > 0$ such that $\|\Omega_n\| \geq D > 0$ for n large enough and

$$CD \leq C \frac{\xi_n^T}{\|\xi_n\|} \Omega_n \frac{\xi_n}{\|\xi_n\|} \leq \frac{\|f(\hat{\theta}_n) - m_n\|}{\|\xi_n\|}.$$

It follows that $\xi_n = O_P(n^{-1/2})$ and that ξ_n is an interior point of $\{\xi \in \mathbb{R}^l \text{ s.t. } \|\xi\| < n^{-1/4}\}$; by concavity of the functional U, ξ_n is the unique maximizer, hence $\xi_n = \hat{\xi}_n$.
We write the first order conditions of optimality of $(\hat{\theta}_n - \theta_0, \hat{\xi}_n)$:

$$\begin{cases} (f(\hat{\theta}_n) - f(\theta_0)) + (f(\theta_0) - m_n) - \int \left[\psi'(\hat{\xi}_n K(F_n(x)) - 1\right] K(F_n(x))dx = 0 \\ J_f(\hat{\theta}_n)\hat{\xi}_n = 0 \end{cases}$$

A mean value expansion (since $\theta_0 \in int(\Theta)$) gives the existence of $\bar{\xi}$ and $\bar{\theta}$ such that $\|\bar{\xi}\| < \|\hat{\xi}_n\|$ and $\|\bar{\theta} - \theta_0\| < \|\hat{\theta}_n - \theta_0\|$ such that

$$\begin{cases} J_f(\bar{\theta})(\theta - \theta_0) + (f(\theta_0) - m_n) - \left[\int \psi''(\bar{\xi} K(F_n(x))K(F_n(x))K(F_n(x))dx\right] \hat{\xi}_n = 0 \\ J_f(\hat{\theta}_n)\hat{\xi}_n = 0 \end{cases}.$$

It holds

$$A_n := \begin{pmatrix} J_f(\bar{\theta}) & -\int \psi''(\bar{\xi} K(F_n(x))K(F_n(x))K(F_n(x))dx \\ 0 & J_f(\hat{\theta}_n) \end{pmatrix} \to_p A := \begin{pmatrix} J_0 & -\Omega \\ 0 & J_0 \end{pmatrix}.$$

By the very definition of A_n,

$$A_n \begin{pmatrix} \hat{\theta}_n - \theta_0 \\ \hat{\xi}_n \end{pmatrix} = \begin{pmatrix} m_n - f(\theta_0) \\ 0 \end{pmatrix}.$$

As Ω is non singular and J_0 has full rank, A is non singular and its inverse is given by

$$A^{-1} = \begin{pmatrix} H & M \\ P & H - H^T \end{pmatrix}.$$

Hence by Lemma 1.4

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_n - \theta_0 \\ \hat{\xi}_n \end{pmatrix} = A_n^{-1} \begin{pmatrix} \sqrt{n}(m_n - f(\theta_0)) \\ 0 \end{pmatrix} \to_d A^{-1} \begin{pmatrix} \mathcal{N}_l(0, \Sigma) \\ 0 \end{pmatrix},$$

which ends the proof.

Chapitre 2

Multivariate quantiles and multivariate L-moments

2.1 Motivations and notations

Univariate L-moments are either expressed as sums of order statistics or as projections of the quantile function onto an orthogonal basis of polynomials in $L_2([0;1],\mathbb{R})$. Both concepts of order statistics and of quantile are specific to dimension one which makes non immediate a generalization to multivariate data.
Let $r \in \mathbb{N}_* := \mathbb{N} \setminus \{0\}$. For an identically distributed sample $X_1, ..., X_r$ on $\mathbb{R}$, we note $X_{1:r} \leq ... \leq X_{r:r}$ its order statistics. It should be noted that $X_{1:r}, ..., X_{r:r}$ are still random variables.
Then, if $\mathbb{E}[|X|] < \infty$, the r-th L-moment is defined by :

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}\mathbb{E}[X_{r-k:r}]. \tag{2.1}$$

If we use F to denote the cumulative distribution function (cdf) and define the quantile function for $t \in [0;1]$ as the generalized inverse of F i.e. $Q(t) = \inf\{x \in \mathbb{R} \text{ s.t. } F(x) > t\}$, this definition can be written :

$$\lambda_r = \int_0^1 Q(t)L_r(t)dt \tag{2.2}$$

where the L_r's are the shifted Legendre polynomials which are a Hilbert orthogonal basis for $L^2([0;1],\mathbb{R})$ equipped with the usual scalar product (for $f, g \in L^2([0;1],\mathbb{R})$, $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$) :

$$L_r(t) = \sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}^2 t^{r-1-k}(1-t)^k = \sum_{k=0}^{r-1}(-1)^{r-k}\binom{r-1}{k}\binom{r-1+k}{k}t^k. \tag{2.3}$$

L-moments were introduced by Hosking [64] in 1990 as alternative descriptors to central moments for a univariate distribution. They have some properties that we wish to keep for the analysis of multivariate data. Serfling and Xiao [99] listed the following key features of univariate L-moments which are desirable for a multivariate generalization :

- The existence of the r-th L-moment for all r if the expectation of the underlying random variable is finite
- A distribution is characterized by its infinite series of L-moments (if the expectation is finite)

- A scalar product representation with mutually orthogonal weight functions (equation 2.2)
- A representation as expected value of an L-statistic (linear function of order statistics)
- The U-statistic structure of sample versions should give asymptotic results
- The L-statistic structure of sample versions should give a quick computation
- Tractable unbiased sample version coming from the U-statistic and L-statistic structure should exist
- Sample L-moments are more stable than classical moments, increasingly with higher order : the impact of each outlier is linear in the L-moment case whereas it is in the order of $(x - \bar{x})^k$ for classical moments of k order

We will add two more properties related to the previous list :

- the equivariance of the L-moments with respect to the dilatation and their invariance with respect to translation for L-moments of an order larger than two
- the tractability of the L-moments in some parametric families which makes them useful for estimation in these families, especially for the shape parameter of heavy tailed distributions.

Heavy-tailed distributions naturally appear in many different fields which then need description features for dispersion or kurtosis usually assuming moments with order larger than two ; for example in applications in climatology based on annual data such as annual maximum rainfall. In [65], Hosking and Wallis successfully applied univariate L-moments for the inference in the so-called regional frequency analysis that have to deal with heavy-tailed distributions. We can mention furthermore financial risk analysis [68] or target detection in radar [110] that are fields in which multivariate heavy-tailed distributions appear.

Serfling and Xiao proposed a multivariate extension of L-moments for a vector $(X_1, ..., X_d)^T$, based on the conditional distribution of X_i given X_j for all $(i, j) \in \{1, ..., d\}^2$. Their definition satisfies most of the properties of the univariate L-moments, but for the characterization of the multivariate distributions by the family of its L-moments. We generalize their approach by a slightly shift in perspective that will allow us to maintain the characterization property in the multivariate case.
Our starting point for a definition of multivariate L-moments is the characterization as orthogonal projection of the quantile onto an orthogonal basis of polynomials defined on $[0; 1]$. It is not difficult to define orthogonal multi-indice polynomials on $[0; 1]^d$ (see Lemma 2.5). It subsequently remains to define a multivariate quantile.
As there is no total order in $\mathbb{R}^d$, there are many different ways to define a multivariate quantile. Serfling made a survey of the existing approaches [97]. Amongst them, we can cite Chaudhuri's spatial quantiles [39], Zuo and Serfling's depth-based quantiles [113] or the generalized quantile process of Einmahl and Mason [49]. In the DOQR (for Depth-Outlyingness-Quantile-Rank) paradigm given by Serfling [98], multivariate quantiles map the ball of center zero and radius 1 $B_d(0, 1)$ into $\mathbb{R}^d$ without specifying the norm underlying the ball. The definition of an orthogonal basis of polynomials is natural only in $[0; 1]^d$, so we consider only the shifted unit ball for the infinite norm in our proposition of multivariate quantile.
The approach of multivariate quantile that has been chosen uses the notion of transport of measure. Indeed, in the univariate case, the quantile maps the uniform measure on $[0; 1]$ onto the distribution of interest. Galichon and Henry [54] for example proposed to keep this basic property in order to define a multivariate quantile as the optimal transport

between the uniform measure on $[0;1]^d$ and the multivariate distribution. We will adopt this definition by relaxing the optimality of the transport. Furthermore, if we consider the Rosenblatt transport [93] in our definition of multivariate L-moments for bivariate random vectors, we match Serfling and Xiao's proposition [99].

We may define a transport $T : \mathbb{R}^d \to \mathbb{R}^d$ between two measures μ and ν defined on $\mathbb{R}^d$.

Definition 2.5. *The pushforward measure of μ through T is the measure denoted by $T\#\mu$ satisfying*

$$T\#\mu(B) = \mu(T^{-1}(B)) \text{ for every Borel subset } B \text{ of } \mathbb{R}^d \tag{2.4}$$

T is said to be a transport map between μ and ν if $T\#\mu = \nu$. In the following, we will call μ the source measure and ν the target measure.

There exist many ways of transporting a measure onto another one. Let us mention for example the transport of Rosenblatt we just mentioned or the transport of Moser [109]. The transport that has received the most attention is undoubtedly optimal transport. Its first formulation goes back to 1781 by Monge. More recently, it was in particular studied by Gangbo, McCann, Villani [108] [109] [79]. In its modern formulation, an optimal transport minimizes a cost function amongst any possible transports.
These transports were used by Easton and McCulloch [47] in order to generalize the Q-Q plots for multivariate data, a graphical tool close to L-moments that especially shows how far two random samples are apart.

However, it is often difficult to have closed forms of the solution of the minimization problem issued from the optimal transport for two arbitrary measures. This is the reason of the following construction of a multivariate quantile.
Let $\mathcal{N}_d$ be the canonical Gaussian measure on $\mathbb{R}^d$. The mapping $Q_0 : [0;1]^d \to \mathbb{R}^d$ defined through

$$Q_0(t_1, ..., t_d) = \begin{pmatrix} \mathcal{N}_1^{-1}(t_1) \\ \vdots \\ \mathcal{N}_1^{-1}(t_d) \end{pmatrix} \tag{2.5}$$

transports the uniform measure $unif$ on $[0;1]^d$ onto $\mathcal{N}_d$ (it is actually an optimal transport for a quadratic cost). This quantile (or transport) provides the reference measure $\mathcal{N}_d$.

Turning back to the extension of the univariate case, consider $\mu = \mathcal{N}_d$ and ν any measure on $\mathbb{R}^d$. With T defined as in 2.4, we may define a transport from the uniform measure on $[0;1]^d$ onto the measure ν on $\mathbb{R}^d$ by

$$Q := T \circ Q_0. \tag{2.6}$$

Q (which is a transport from $unif$ to ν) is a natural extension of the quantile function defined from $[0;1]$ equipped with the uniform measure onto $\mathbb{R}$ equipped with a given measure.
Clearly, the intermediate Gaussian measure can be skipped and a quantile may be defined directly from $[0;1]^d$ onto $\mathbb{R}^d$ with the respective measures $unif$ and ν.Indeed, we will define transports from $[0;1]^d$ equipped with $unif$ onto $[0;1]^d$ equipped with a given copula ; see Section 2.4.2.
The interest in the intermediate (or reference) Gaussian measure μ lies in the fact that a transport T from μ onto a measure ν will be easy to define when ν belongs to specific

classes of multivariate distributions with rotational parameters. Note that the transport T need not be optimal for some cost.

We will concentrate our attention on models close to elliptical distributions. Let us recall that elliptical distributions are parametrized by the existence of a scatter matrix Σ, a location vector m and a radial scalar random variable $R \in \mathbb{R}_+$. In fact, $X \in \mathbb{R}^d$ follows an elliptical distribution if and only if

$$X \overset{d}{=} m + R\Sigma^{1/2}U$$

with U uniform over $S_{d-1}(0,1)$, the sphere of center zero and radius 1 and R independent of U.
Even if, to our knowledge, there are no tractable closed forms for the optimal transport of the uniform on $[0;1]^d$ (or even of the standard Gaussian) onto an elliptical distribution, we can define a family of models close to the elliptical ones that contains spherical distributions with an explicit quantile. This allows to build estimators based on a multivariate method of L-moments for the scatter matrix and the mean parameters of this family.
The price to pay for using optimal transports is to consider models adapted to this approach. A natural way to work with such quantiles is then to define models through their quantile function, instead of the classical density function. Sei proposed [96] to define models through their transport onto a standard multivariate Gaussian. Such models have desirable properties, in particular the ease to describe the independence of marginals and the concavity of their log-likelihood. In a similar desire to define non-Gaussian distributions easy to manipulate in the context of linear models, Box and Cox used a particular form of this transport as well [26].

Let us now introduce some notation. In the following, we will consider a random variable or vector X with measure ν and $\overset{d}{=}$ means the equality in distribution. The scalar product between x and y in $\mathbb{R}^d$ will be noted $x.y$ or $\langle x, y\rangle$.

2.2 Definition of multivariate L-moments and examples

2.2.1 General definition of multivariate L-moments

Let X be a random vector in $\mathbb{R}^d$. We wish to exploit the representation given by the equation (2.2) in order to define multivariate L-moments. Recall that we chose quantiles as mappings between $[0;1]^d$ and $\mathbb{R}^d$.

We explicit a polynomial orthogonal basis on $[0;1]^d$. Let $\alpha = (i_1, ..., i_d) \in \mathbb{N}^d$ be a multi-index and $L_\alpha(t_1, ..., t_d) = \prod_{k=1}^d L_{i_k}(t_k)$ (where the L_{i_k}'s are univariate Legendre polynomials defined by equation 2.3) the natural multivariate extension of the Legendre polynomials. Indeed, it holds

Lemma 2.5. *The L_α family is orthogonal and complete in the Hilbert space $L^2([0;1]^d, \mathbb{R})$ equipped with the usual scalar product :*

$$\forall f, g \in L^2([0;1]^d), \quad \langle f, g\rangle = \int_{[0;1]^d} f(u).g(u)du \tag{2.7}$$

Proof. The orthogonality is straightforward since if $\alpha = (i_1, ..., i_d) \neq \alpha' = (i'_1, ..., i'_d)$, there exists a subindex $1 \leq k \leq d$ such that $i_k \neq i'_k$ and

$$\int_{[0;1]^d} L_\alpha(t_1, ..., t_d) L_{\alpha'}(t_1, ..., t_d) dt_1 ... dt_d = \prod_{j=1}^{d} \int_0^1 L_{i_j}(t_j) L_{i'_j}(t_j) dt_j = 0 \tag{2.8}$$

thanks to the orthogonality of $L_{i'_k}$ and L_{i_k} in $L_2([0;1], \mathbb{R})$.
The univariate Legendre polynomials define an orthogonal basis for the space of polynomials denoted by $\mathbb{R}[X]$. Hence, for all k, there exists $c_1, ..., c_k \in \mathbb{R}$ such that $X^k = \sum_{i=1}^{k} c_i L_i(X)$. Thus for all $k_1, ..., k_d$, there exists $c_{11}, ..., c_{1k}, ..., c_{d1}, ..., c_{dk} \in \mathbb{R}$ such that

$$\prod_{j=1}^{d} X_j^{k_j} = \prod_{j=1}^{d} \left(\sum_{i=1}^{k_j} c_{ji} L_i(X) \right).$$

We deduce that (L_α) is an orthogonal basis of the space of polynomial with d indices $\mathbb{R}[X_1, ..., X_d]$. It remains to prove that $\mathbb{R}[X_1, ..., X_d]$ is dense in $L_2([0;1]^d, \mathbb{R})$.
For this purpose, let $f \in L_2([0;1]^d, \mathbb{R})$. We define a test function $\varphi \in C^0([0;1]^d, \mathbb{R})$ defined for $x \in [0;1]^d$

$$\varphi(x) = \begin{cases} e^{-\frac{1}{1-\|x\|^2}} & \text{if } \|x\| < 1 \\ 0 & \text{if } \|x\| = 1 \end{cases}$$

with $\|x\| = \sqrt{\sum_{i=1}^{d} x_i^2}$.
Let n be an integer greater than zero and

$$f_n(x) = \frac{1}{\int_{\mathbb{R}^d} \varphi(x) dx} \int_{[\mathbb{R}^d} \frac{1}{n^d} f(x-y) \varphi(\frac{y}{n}) \mathbb{1}_{x-y \in [0;1]^d} dy$$

Then for all $n > 0, f_n \in C^0([0;1]^d, \mathbb{R}^d)$ and $f_n \to f$ in $L_2([0;1]^d, \mathbb{R})$. Indeed, by noting $a = \int_{\mathbb{R}^d} \varphi(x) dx$ for $x \in [0;1]^d$

$$\begin{aligned} f_n(x) - f(x) &= \frac{1}{a} \int_{[\mathbb{R}^d} (f(x-y) - f(x)) \frac{1}{n^d} \varphi(\frac{y}{n}) \mathbb{1}_{x-y \in [0;1]^d} dy \\ &= \frac{1}{a} \int_{[\mathbb{R}^d} (f(x-ny) - f(x)) \varphi(y) \mathbb{1}_{x-ny \in [0;1]^d} dy \end{aligned}$$

Furthermore

$$\|f(x-ny) - f(x)\|^2 \mathbb{1}_{x-ny \in [0;1]^d} \varphi(y)^2 \leq 2\varphi(y)^2 \int_{[0;1]^d} f(y)^2 dy = \|f\|_{L_2}^2 \varphi(y)^2,$$

Then as for any $y \in \mathbb{R}^d$, $\|f(x-ny) - f(x)\|^2 \mathbb{1}_{x-ny \in [0;1]^d} \to 0$ when $n \to \infty$; we apply the dominated convergence theorem to show that $\|f_n(x) - f(x)\|^2 \to 0$ for any $x \in [0;1]^d$. In the same way, as

$$\|f_n(x) - f(x)\|^2 \leq 2 \int_{[0;1]^d} f(y)^2 dy$$

We prove by a second application of the dominated convergence theorem that $f_n \to f$ in $L_2([0;1]^d, \mathbb{R})$.
Let $\epsilon > 0$. We can thus find $N > 0$ such that

$$\|f - f_N\|_{L_2} < \epsilon$$

Hence, as $f_N \in C^0([0;1]^d, \mathbb{R}^d)$, by Stone-Weierstrass Theorem (see for example Rudin Theorem 5.8 [94]), there exists $g \in \mathbb{R}[X_1, ..., X_d]$ such that :

$$\|f_N - g\|_\infty < \epsilon.$$

Then $\|f - g\|_{L_2} < \|f - f_N\|_{L_2} + \|f_N - g\|_{L_2} < \epsilon + \|f_N - g\|_\infty < 2\epsilon$. We conclude that $\mathbb{R}[X_1, ..., X_d]$ is dense in $L_2([0;1]^d, \mathbb{R})$ which proves that $(L_\alpha)_{\alpha \in \mathbb{N}_*^d}$ is complete. □

We can finally define the multivariate L-moments.

Definition 2.6. *Let $Q : [0;1]^d \to \mathbb{R}^d$ be a transport between the uniform distribution on $[0;1]^d$ and ν. Then, if $\mathbb{E}[\|X\|] < \infty$, the L-moment λ_α of multi-index α associated to the transport Q are defined by :*

$$\lambda_\alpha := \int_{[0;1]^d} Q(t_1, ..., t_d) L_\alpha(t_1, ..., t_d) dt_1 ... dt_d \in \mathbb{R}^d. \tag{2.9}$$

With this definition, there are as many L-moments as ways to transport $unif$ onto ν. The hypothesis of finite expectation guarantees the existence of all L-moments :

$$\begin{aligned} \left\| \int_{[0;1]^d} Q(t_1, ..., t_d) L_\alpha(t_1, ..., t_d) dt_1 ... dt_d \right\| &\leq \left(\sup_{t \in [0;1]^d} |L_\alpha(t)| \right) \int_{[0;1]^d} \|Q(t_1, ..., t_d)\| \, dt_1 ... dt_d \\ &\leq \int_{[0;1]^d} \|x\| dF(x) < \infty. \end{aligned}$$

Remark 2.19. *Given the degree δ of $\alpha = (i_1, ..., i_d)$ that we define by $\delta = \sum_{k=1}^d (i_k - 1) + 1$, we may define all L-moments with degree δ, each one associated with a given corresponding α leading to the same δ.*
For example, the L-moment of degree 1 is

$$\lambda_1 (= \lambda_{1,1,...,1}) = \int_{[0;1]^d} Q(t_1, ..., t_d) dt_1 ... dt_d = \mathbb{E}[X]. \tag{2.10}$$

The L-moments of degree 2 can be grouped in a matrix :

$$\Lambda_2 = \left[\int_{[0;1]^d} Q_i(t_1, ..., t_d)(2t_j - 1) dt_1 ... dt_d \right]_{1 \leq i,j \leq d}. \tag{2.11}$$

In equation 2.10 we noted $Q(t_1, ..., t_d) = \begin{pmatrix} Q_1(t_1, ..., t_d) \\ \vdots \\ Q_d(t_1, ..., t_d) \end{pmatrix}$.

Proposition 2.7. *Let ν and ν' be two Borel probability measures. We suppose that Q and Q' respectively transport $unif$ onto ν and ν'.*
Assume that Q and Q' have same multivariate L-moments $(\lambda_\alpha)_{\alpha \in \mathbb{N}_*^d}$ given by the equation (2.9).
Then $\nu = \nu'$. Moreover :

$$Q(t_1, ..., t_d) = \sum_{(i_1,...,i_d) \in \mathbb{N}_*^d} \left(\prod_{k=1}^d (2i_k + 1) \right) L_{(i_1,...,i_d)}(t_1, ..., t_d) \lambda_{(i_1,...,i_d)} \in \mathbb{R}^d \tag{2.12}$$

Proof. We have to prove that if Q and Q' are two transports coming from ν and ν' such that all their L-moments coincide, $\nu = \nu'$.
We denote by λ_α and λ'_α their respective L-moments of multi-index α.
As the Legendre family is orthogonal and complete in $L_2([0;1]^d, \mathbb{R})$, we can decompose each component of Q :

$$\begin{aligned} Q(t_1, ..., t_d) &= \sum_{\alpha \in \mathbb{N}_d} \frac{\langle Q, L_\alpha \rangle_{L_2}}{\langle L_\alpha, L_\alpha \rangle_{L_2}} L_\alpha(t_1, ..., t_d) \\ &= \sum_{\alpha \in \mathbb{N}_*^d} \left(\prod_{k=1}^{d} (2i_k + 1) \right) \lambda_\alpha L_\alpha(t_1, ..., t_d) \end{aligned}$$

because for $\alpha = (i_1, ..., i_d) \in \mathbb{N}_*^d$

$$\int_{[0;1]^d} L_\alpha(t_1, ..., t_d)^2 dt_1 ... dt_d = \prod_{k=1}^{d} ||L_{i_k}||^2_{L_2([0;1])} = \prod_{k=1}^{d} \frac{1}{2i_k + 1}.$$

By the same reasoning, we get

$$Q'(t_1, ..., t_d) = \sum_{\alpha \in \mathbb{N}_*^d} \left(\prod_{k=1}^{d} (2i_k + 1) \right) \lambda'_\alpha L_\alpha(t_1, ..., t_d).$$

We conclude that $Q = Q'$ and $\nu = \nu'$ by hypothesis. □

2.2.2 L-moments ratios

Let us note $\lambda_r(X)$ the r-th univariate L-moment of the random variable X and $(b_1, ..., b_d)$ the canonical basis of $\mathbb{R}^d$. Let us decompose the vector λ_α into

$$\lambda_\alpha = \begin{pmatrix} \langle \lambda_\alpha, b_1 \rangle \\ \vdots \\ \langle \lambda_\alpha, b_d \rangle \end{pmatrix} \in \mathbb{R}^d.$$

Definition 2.7. *As for univariate L-moments, we can define normalized ratios of L-moments for any multi-index $\alpha \in \mathbb{N}^d$ different from (1,...,1) by :*

$$\tau_\alpha = \begin{pmatrix} \langle \tau_\alpha, b_1 \rangle \\ \vdots \\ \langle \tau_\alpha, b_d \rangle \end{pmatrix} = \begin{pmatrix} \frac{\langle \lambda_\alpha, b_1 \rangle}{\lambda_2(X_1)} \\ \vdots \\ \frac{\langle \lambda_\alpha, b_d \rangle}{\lambda_2(X_d)} \end{pmatrix}. \tag{2.13}$$

with $\lambda_2(X_i)$ denoting the univariate second L-moment related to X_i.

This definition is guided by the following inequality :

Proposition 2.8. *For all $\alpha \in \mathbb{N}_*^d$ different from (1,...,1), we have :*

$$|\langle \tau_\alpha, e_i \rangle| \leq 2; \tag{2.14}$$

Moreover, if $\alpha = (i_1, ..., i_d)$ with $i_j = 2$ and $i_k = 1$ for all $k \neq j$, let $U = (U_1, ..., U_d)^T$ be a uniform random vector on $[0;1]^d$ and $U_{-j} = (U_1, ..., U_{j-1}, U_{j+1}, ..., U_d)^T$ and $V = \mathbb{E}_{U_{-j}}[Q_i(U)]$.
Then

$$|\langle \tau_\alpha, b_i \rangle| \leq \frac{\lambda_2(V)}{\lambda_2(X_i)} \tag{2.15}$$

Proof. Let $y \in \mathbb{R}$. Then as $\alpha \neq (1, \dots, 1)$,

$$\begin{aligned}\langle \lambda_\alpha, b_i \rangle &= \int_{[0;1]^d} Q_i(t_1, ..., t_d) L_\alpha(t) dt_1 ... dt_d \\ &= \int_{[0;1]^d} (Q_i(t_1, ..., t_d) - y) L_\alpha(t) dt_1 ... dt_d.\end{aligned}$$

As $|L_\alpha(t)| \leq 1$ for all $t \in [0;1]^d$, by definition of the transport, it holds

$$\begin{aligned}|\langle \lambda_\alpha, b_i \rangle| &\leq \int_{[0;1]^d} |Q_i(t_1, ..., t_d) - y| \, dt_1 ... dt_d \\ &\leq \int_0^1 \int_0^1 |x_i - y| \, dF_i(x_i) dF_i(y) \\ &\leq \mathbb{E}_{X_i \overset{d}{=} Y_i}[|X_i - Y_i|] = 2\lambda_2(X_i).\end{aligned}$$

This proves the first assertion. The second is inspired from the proposition 4 of [99]. As the degree of α is 2, there exists $1 \leq j \leq d$ such that

$$\langle \lambda_\alpha, b_i \rangle = \int \left(\int Q_i(t_1, ..., t_d) L_2(t_j) dt_j \right) dt_1 ... dt_{j-1} dt_{j+1} ... dt_d.$$

We note $U_{-i} = (U_1, ..., U_{j-1}, U_{j+1}, ..., U_d)'$, $V = \mathbb{E}_{U_{-j}}[Q_i(U)]$ and $W = U_j$. Then by noting

$$\begin{aligned}\langle \lambda_\alpha, b_i \rangle &= \mathbb{E}[V L_2(W)] \\ &= 2\mathbb{E}[VW] - \mathbb{E}[V] \\ &= 2Cov(V, W)\end{aligned}$$

where V and W are two random variables of finite expectation and covariance. Then, Hoeffding lemma quoted in [75] gives us :

$$Cov(V, W) = \int \int [F_{V,W}(v, w) - F_V(v) F_W(w)] \, dv dw$$

Moreover, the well-known Fréchet bounds assert that for any v, w

$$\max(F_W(w) + F_V(v) - 1, 0) \leq F_{v,W}(v, w) \leq \min(F_V(v), F_W(w)).$$

Since W is uniform on $[0;1]$

$$Cov(V, W) \leq \int \int [\min(F_V(v), w) - F_V(v) w] \, dv dw.$$

Furthermore

$$Cov(V, F_V(V)) = \int \int [\min(F_V(v), w) - F_V(v) w] \, dv dw.$$

We conclude that

$$Cov(V, W) \leq Cov(V, F_V(V)).$$

Now, using $\max(a + b - 1, 0) - ab = -(\min(1 - a, b) - (1 - a)b)$ along with the Fréchet bound, a similar reasoning leads to

$$Cov(V, W) \geq -Cov(V, F_V(V)).$$

Remarking that $2Cov(V, F_V(V)) = \lambda_2(V)$, we obtain

$$|\langle \lambda_\alpha, b_i \rangle| \leq \lambda_2(V).$$

□

Remark 2.20. *The inequality in the previous Proposition is probably not optimal but has the advantage of some generality. As we will see later, if we choose the particular bivariate Rosenblatt transport, it holds* $|\langle \tau_\alpha, b_i \rangle| \leq 1$ *for* $\alpha = (1, 2)$ *or* $\alpha = (2, 1)$.

2.2.3 Compatibility with univariate L-moments

The definition which we adopted for the definition of general L-moments is compatible with the similar one in dimension 1 since the univariate quantile is a transport.

Definition 2.8. *Let ν be a real probability measure. The quantile is the generalized inverse of the distribution function :*

$$Q(t) = \inf\{x \in \mathbb{R} \text{ s.t. } \nu((-\infty; x]) \geq t\}. \tag{2.16}$$

Proposition 2.9. *If we denote by μ the uniform measure on $[0;1]$, then $Q\#\mu = \nu$ i.e. $Q(U) \overset{d}{=} X$ if U denotes the uniform law on $[0;1]$, and X denotes the random variable associated to ν.*

Proof. Let $x \in \mathbb{R}$. We denote by F the cdf of X and by A_t the event

$$A_t = \{x \in \mathbb{R} \text{ s.t. } F(x) \geq t\}$$

We then have $Q(t) = \inf A_t$. We wish to prove :

$$\{t \in [0;1] \text{ s.t. } Q(t) \leq x\} = \{t \in [0;1] \text{ s.t. } t \leq F(x)\} \tag{2.17}$$

We temporarily admit this assertion. Then

$$\begin{aligned} \mathbb{P}[Q(U) \leq x] &= \mathbb{P}[U \leq F(x)] \\ &= F(x) \end{aligned}$$

which ends the proof. It remains to prove 2.17.
First, the definition of Q gives us

$$\{t \leq F(x)\} \Rightarrow \{x \in A_t\} \Rightarrow \{Q(t) \leq x\}$$

Secondly, let t be such that $Q(t) \leq x$. Then by monotony of F, $F(Q(t)) \leq F(x)$. We then claim that

$$Q(t) \in A_t$$

Indeed, let us suppose the contrary and consider a strictly decreasing sequence $x_n \in A_t$ such that

$$\lim_{n\to\infty} x_n = \inf A_t = Q(t).$$

By right continuity of F

$$\lim_{n\to\infty} F(x_n) = F(Q(t))$$

and, on the other hand, by definition of A_t,

$$\lim_{n\to\infty} F(x_n) \geq t$$

i.e. $Q(t) \in A_t$ wihch contradicts the hypothesis. Then $Q(t) \in A_t$ i.e. $t \leq F(Q(t))$ thus $t \leq F(x)$. We have proved that

$$\{Q(t) \leq x\} \Rightarrow \{t \leq F(x)\}$$

□

Subsequently, if we consider the particular transport defined by the univariate quantile, the L-moments are defined by

$$\lambda_r = \int_0^1 Q(t) L_r(t) dt \tag{2.18}$$

which is the quantile characterization of univariate L-moments.

Remark 2.21. *This transport corresponds to a Rosenblatt transport and an optimal transport with respect to a large family of costs (see Proposition 2.11).*

2.2.4 Relation with depth-based quantiles

With the DOQR paradigm, Sefling related the four following notions :

- the **centered quantile function** : a centered multivariate quantile function Q indexed by $u \in B_d$, the unit ball in $\mathbb{R}^d$ such that $x := Q(u)$ is a centered quantile representation of x. $Q(0)$ represents the center of mass or median. This quantile function generates nested contours $\{Q(u) : \|u\| = c\}$ grouping points of the distribution by "distance" to the center of mass.
- the **centered rank function** : if the quantile $Q : B_d \to \mathbb{R}^d$ has an inverse, noted $R : \mathbb{R}^d \to B_d$, it corresponds to the centered rank function. For each point x, $R(x)$ corresponds to the directional rank of x.
- The **outlyingness function** : the magnitude $O(x) := \|R(x)\|$ defines a measure of the outlyingness of x.
- The **depth function** : the magnitude $D(x) := 1 - O(x)$ provides a center-outward ordering of x, higher depth corresponding to higher centrality.

With this paradigm, all the depth functions introduced for example in [113] can induce a quantile function (see [98]). Even if the quantile deduced from a depth function is not uniquely defined, the contours associated to the depth are unique.

If we note Q the quantile as a transport between the uniform distribution in $[0;1]^d$ and the distribution of interest, then the function

$$\tilde{Q} := u \in [-1;1]^d \mapsto Q(\frac{u}{2} - (1/2, ..., 1/2)^T)$$

correspond to the Serfling's notion of centered quantile for the infinite norm. If Q is invertible, we can therefore introduce a related depth function as

$$D(x) = 1 - 2\|Q^{-1}(x) - (1/2, ..., 1/2)^T\|.$$

This allows us to compare this depth function with respect to the desirable criteria for a depth function enounced in [113] satisfied by classical depth functions such as Tukey's half-space depth function.

- **Affine invariance** : the depth of a point $x \in \mathbb{R}^d$ should not depend on the underlying coordinate system. This property is not verified by the depth issued from transport and should be a stake for future works.
- **Maximality at center** : the obvious center $Q(1/2, ..., 1/2)$ is the point of maximal depth
- **Monotonicity relative to deepest point** : as the point $x \in \mathbb{R}^d$ moves away from the center of mass, the depth function evaluated on x decreases monotonically. This intuitive property should restrict the transports acceptable for Q to be a quantile. For monotone and Rosenblatt transports introduced in the sequel, this property holds.
- **Vanishing at infinity** : the depth of a point x should approach zero as $\|x\|$ approaches infinity.

The quantile function issued from a transport brings moreover indications on the location of the mass of the multivariate distribution of measure ν. Indeed, all intuitive information of a "piece" of the unit cube (centrality, extremality, volume,...) can be transposable to the transported piece of points in $\mathbb{R}^d$. In mathematical terms, if A is Borelian of $[0;1]^d$, it holds :

$$\nu(Q(A)) = \mu(A) = vol(A)$$

We will now consider in the following two different kinds of transport among many others :

- the optimal transport
- the Rosenblatt transport

2.3 Optimal transport

2.3.1 Formulation of the problem and main results

Let us consider two measures μ and ν respectively defined on $\Omega \subset \mathbb{R}^d$ and $\mathbb{R}^d$. If we define a cost function $c : \Omega \times \Omega \to \mathbb{R}$, then the problem is to find an application T that transports μ into ν and minimizes :

$$\int_\Omega c(x, T(x)) d\mu(x). \tag{2.19}$$

The quadratic case $c : (x, y) \mapsto (x - y)^2$ was first studied by Brenier [27], the generalization to generic costs has been considered, among others, by McCann, Gangbo, Villani [79][108]. Let us give the following theorem for specific convex costs $(x, y) \mapsto c(x, y) = h(x - y)$:

Theorem 2.4. *(McCann, Gangbo)*
Let $h : \mathbb{R}^d \to \mathbb{R}$ be a convex function, μ and ν be two probability measures on $\mathbb{R}^d$. Let us suppose that there exists a transport T such that $\int_{\mathbb{R}^d} h(x - T(x)) d\mu(x) < \infty$. Let us assume that μ is absolutely continuous with respect to the Lebesgue measure.
Then, there exists a unique transport T from μ to ν that minimizes the cost $\int_\Omega h(x - T(x)) d\mu(x)$ determined $d\mu$-almost everywhere and characterized by a function ϕ :

$$T(x) = x - \nabla h^* (\nabla \phi(x)) \tag{2.20}$$

where h^* is the Legendre transform of h.

$$h^*(y) = \sup_{x \in \mathbb{R}^d} \langle x, y \rangle - h(x).$$

The function ϕ is $d\mu$-a.s. unique up to an arbitrary additive constant.

Proof. We apply Theorem 2.44 of [108]. □

Remark 2.22. *If we consider the quadratic case $h(x - y) = (x - y)^2$, the above existence theorem is equivalent to the existence of another function (that will be called potential function) $\varphi := x \mapsto \|x\|^2 - \phi(x)$ which is convex such that the optimal transport is $T = \nabla \varphi$. We can observe a refinement of this case in the following Proposition 2.10.*

For the definition of a multivariate quantile, μ is the uniform measure on $\Omega = [0; 1]^d$ and ν is the measure of a random vector X of interest. The corresponding transport will be denoted by Q.
As μ is absolutely continuous with respect to the Lebesgue measure, the remaining assumption in Theorem 2.4 is the existence of a transport Q such that $Q \# \mu = \nu$ and $\int_\Omega h(Q(u) - u) du < \infty$.
We can remove this limitation by considering source measures μ that give no mass to "small sets". To make the term "small set" more precise, we use the Hausdorff dimension.

Definition 2.9. *Let E be a metric space. If $S \subset E$ and $p \geq 0$, the p-dimensional Hausdorff content of S is defined by :*

$$C_d(S) = \inf \left\{ \sum_i r_i^p \text{ such that there is a cover of } S \text{ by balls with radii } r_i > 0 \right\}.$$

Then, the Hausdorff dimension of E is given by :

$$dim(E) := \inf \{d \geq 0 \text{ such that } C_d(E) = 0\} \tag{2.21}$$

Proposition 2.10. *(McCann/Brenier's Theorem)*
Let μ, ν be two probability measures on $\mathbb{R}^d$, such that μ does not give mass to sets of Hausdorff dimension at most $d-1$. Then, there is exactly one measurable map T such that $T\#\mu = \nu$ and $T = \nabla\varphi$ for some convex function φ, in the sense that any two such maps coincide $d\mu$-almost everywhere.

Proof. Theorem 2.32 of [108] □

Remark 2.23. *When μ is the uniform measure on $[0;1]^d$, Proposition 2.10 holds for any ν.*

The gradient of convex potentials are called monotone by analogy with the univariate case. We can see this gradient as the solution of a potential differential equation. By abuse of language, we will refer at this transport as monotone transport in the sequel.

Remark 2.24. *Let us suppose μ and ν admit densities with respect to the Lebesgue measure respectively denoted by p and q. Proposition 2.10 provides a mapping $\nabla\varphi$ such that for all test functions a in C^∞ with a compact support :*

$$\int a(y)q(y)dy = \int a(\nabla\varphi(x))p(x)dx.$$

Let us assume furthermore that $\nabla\varphi$ is C^1 and bijective. We can then perform the change of variables $y = \nabla\varphi(x)$ on the left-hand side of the previous equality :

$$\int a(y)q(y)dy = \int a(\nabla\varphi(x))q(\nabla\varphi(x))\det(\nabla^2\varphi(x))dx.$$

Since the function a is arbitrary, we get :

$$p(x) = q(\nabla\phi(x)))\det(\nabla^2\phi(x)). \tag{2.22}$$

This is a particular case of the general Monge-Ampère equation

$$\det(\nabla^2\varphi(x)) = F(x, \varphi(x), \nabla\varphi(x)).$$

2.3.2 Optimal transport in dimension 1

The natural order of the real line implies that the quantile is a solution of several transport problems :

Proposition 2.11. *(Optimal transport in dimension $d = 1$)*
Let μ and ν be two arbitrary measures respectively defined on $[0;1]$ and $\mathbb{R}$ such that μ gives no mass to atoms. Let $T : [0;1] \to \mathbb{R}$ be a transport of μ onto ν. Then for any real convex function h :

$$\int_0^1 h(Q(F(u)) - u)du \leq \int_0^1 h(T(u) - u))du \tag{2.23}$$

where Q is the generalized quantile of ν and F the cdf of μ i.e. $Q \circ F$ is the solution of the univariate transport problem.

Proof. We refer to Theorem 6.0 [4]. □

This result is particular to the dimension 1. If we plug this result in the definition we gave for multivariate L-moments applied with $d = 1$, we obtain the univariate definition of L-moments :

$$\lambda_r = \int_0^1 Q(t)L_r(t)dt. \tag{2.24}$$

The multivariate L-moments defined with the optimal transport are then compatible with the definition in dimension $d = 1$.

2.3.3 Examples of monotone transports

Example 2.14. *(Univariate Gaussian)*
Let us consider the univariate Gaussian $\mathcal{N}_{m,\sigma}$ with $m \in \mathbb{R}$ and $\sigma > 0$. The potential is then defined up to a constant by :

$$\text{for } t \in [0;1], \quad \phi_{m,\sigma}(t) = \int_{1/2}^{t} \left(m + \sigma\mathcal{N}^{-1}(u)\right) du \tag{2.25}$$

where $\mathcal{N}$ is the cumulative distribution function of the standard Gaussian.

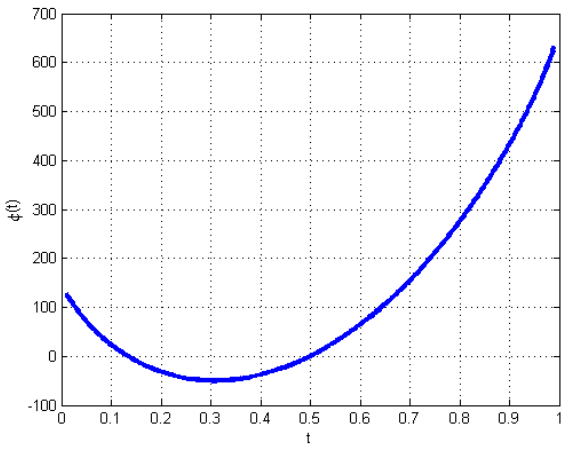


Figure 2.1 – Potential function $\phi_{m,\sigma}$ for $m = 1$ $\sigma = 2$

Let us note that the potential is minimum at the cumulative weight t such that $\mathcal{N}_{m,\sigma}^{-1}(t) = 0$ and is equal to 0 at the median.
The gradient is simply the quantile

$$\nabla\phi_{m,\sigma}(t) = m + \sigma\mathcal{N}^{-1}(t) = Q_{\mathcal{N}}(t).$$

If we build the Legendre transform of the potential, we find a dual potential :

$$\psi_{m,\sigma}(x) = \sup_{t\in[0;1]} \{xt - \phi_{m,\sigma}(t)\} = x\mathcal{N}\left(\frac{x-m}{\sigma}\right) - \frac{1}{\sigma}\int_m^x x\mathcal{N}'\left(\frac{x-m}{\sigma}\right) dy,$$

and

$$\nabla\psi_{m,\sigma}(x) = \mathcal{N}\left(\frac{x-m}{\sigma}\right).$$

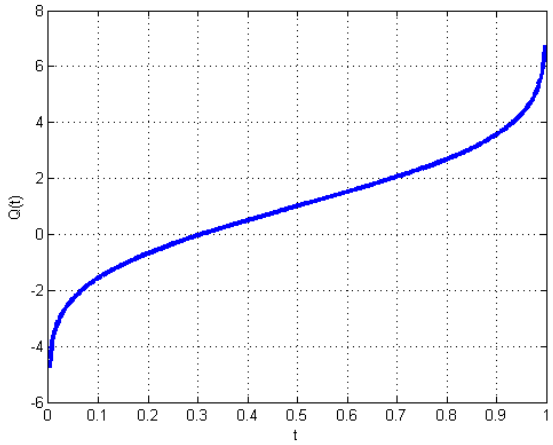


FIGURE 2.2 – Quantile function $\nabla\phi_{m,\sigma}$ for $m = 1$ $\sigma = 2$

Finally the Hessian of the dual potential is the density of the Gaussian as expected by the Monge-Ampère equation

$$\nabla^2\psi_{m,\sigma}(x) = \frac{1}{\sigma}\mathcal{N}'\left(\frac{x-m}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

Example 2.15. *(Independent coordinates)*
For a vector of independent marginals $(X_1, ..., X_d)$, the optimal transport is easily obtained since it is the concatenation of each marginal univariate quantile :

$$Q(t_1, ..., t_d) = \begin{pmatrix} Q_1(t_1) \\ \vdots \\ Q_d(t_d) \end{pmatrix} \tag{2.26}$$

if $Q_1, ..., Q_d$ are the respective quantiles of $X_1, ..., X_d$.
Indeed, it is obvious that the mapping Q defined above transports the uniform measure on $[0;1]^d$ into the distribution of $(X_1, ..., X_d)$.
Furthermore, as $Q_1, ..., Q_d$ are univariate transports, they are gradients of convex functions that can be denoted by respective potentials $\phi_1, ..., \phi_d : [0;1] \to \mathbb{R}$. Then, if we build the potential :

$$\phi(t_1, ..., t_d) = \phi_1(t_1) + \cdots + \phi_d(t_d), \tag{2.27}$$

we remark that $\nabla\phi = Q$ and ϕ is convex because each ϕ_i is convex.

Example 2.16. *(Max-Copula)*
Let us define a 2-dimensional potential for $u, v \in [0;1]^2$:

$$\phi(u, v) = \frac{1}{4}(u + v)^2. \tag{2.28}$$

ϕ is convex (but not strictly convex) and derivable almost everywhere ; the associated transport is

$$T(u, v) = \nabla\phi(u, v) = \begin{pmatrix} \frac{u+v}{2} \\ \frac{u+v}{2} \end{pmatrix}$$

T then transports the uniform distribution on $[0;1]^2$ into the distribution defined by the cdf $F(u, v) = \min(u, v)$.

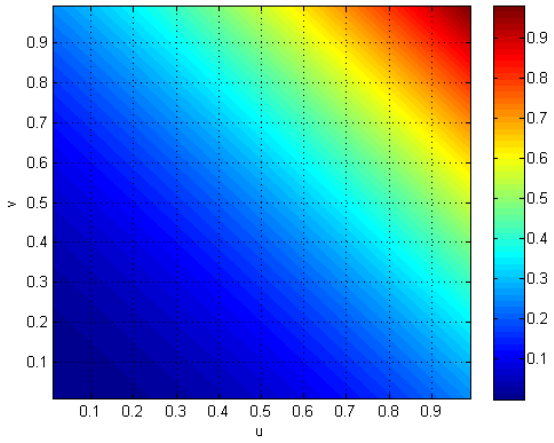


FIGURE 2.3 – Color levels of the potential function ϕ

The distribution considered in the previous example corresponds to the max-copula. A copula induces a distribution defined on $[0;1]^2$ with uniform margins and is a measure of the dependence for a bivariate random vector.

2.4 L-moments issued from the monotone transport

From now on, the notion of optimal transport will uniquely refer to the monotone case.

2.4.1 Monotone transport from the uniform distribution on $[0;1]^d$

Let $X \in \mathbb{R}^d$ be a random vector. According to Brenier's Theorem, there exists a potential $\varphi : [0;1]^d \to \mathbb{R}^d$ such that

$$\nabla\varphi(U) \overset{d}{=} X.$$

The α-th multivariate L-moment associated to this potential is

$$\lambda_\alpha = \int_{[0;1]^d} \nabla\varphi(t) L_\alpha(t) dt. \tag{2.29}$$

We keep the property of invariance with respect to translation and equivariance with respect to dilatation coming from the univariate L-moments

Proposition 2.12. *Let X be a random vector in $\mathbb{R}^d$ and $\lambda_\alpha(X)$ its associated L-moments such that*

$$\lambda_\alpha(X) = \int_{[0;1]^d} \nabla\varphi(t) L_\alpha(t) dt \tag{2.30}$$

with $\nabla\varphi$ the transport from the uniform on $[0;1]^d$ onto X and φ convex.
Let $m \in \mathbb{R}^d$ and $\sigma > 0$. Then

$$\lambda_\alpha(m + \sigma X) = \sigma\lambda_\alpha + m\mathbb{1}_{\alpha=(1\ldots1)} \tag{2.31}$$

Proof. Let $\psi : x \mapsto \sigma\varphi(x) + \langle x, m\rangle$. Then ψ is convex and $\nabla\psi(X) = \sigma X + m$.
$\nabla\psi$ is then the monotone transport from the uniform distribution on $[0;1]^d$ onto the distribution of $\sigma X + m$. □

We do not have a general property dealing with rotation with this transport. This is strongly due to the bad behavior of the unit square through this transformation. We will present later Hermite L-moments that partially fill this deficiency.

2.4.2 Monotone transport for copulas

Let X be a bivariate vector of cdf denoted by H. We can build a transport of the bivariate uniform distribution on $[0;1]^2$ into X through the composition of the transport of the copula of X with the transport of the marginals. The reason of this construction is that the copula function is well adapted to the unit square $[0;1]^d$.

Let us first present the definition of a copula and Sklar's Theorem.

Definition 2.10. *A copula is a function $C : [0;1]^2 \to [0;1]$ with the following properties :*

- *C is 2-increasing i.e. for all $u_1 \leq u_2 \in [0;1]$ and $v_1 \leq v_2 \in [0;1]$:*

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$$

- *for $u, v \in [0;1]$:*

$$C(u, 1) = u \ , \ C(u, 0) = 0 \text{ and } C(1, v) = v \ , \ C(0, v) = 0$$

Theorem 2.5. *(Sklar's theorem)*
Let H be a joint distribution with margins F and G. Then there exists a copula C such that for all $x, y \in \bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$:

$$H(x, y) = C(F(x), G(y)) \tag{2.32}$$

C is uniquely defined on $F(\bar{\mathbb{R}}) \times G(\bar{\mathbb{R}})$.
Conversely, if C is a copula and F and G are distribution functions, then H defined by the above equation is a joint distribution function with margins F and G.

Proof. see Theorem 2.3.3 of Nelsen [81]. □

Let now C be a copula associated to the bivariate vector X. Then, using the previous Theorem 2.5, C is the joint distribution function of a bivariate vector W with uniform margins on $[0;1]$. Let Q_C be the optimal transport of U with uniform distribution on $[0;1]^2$ into $W = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$. As W and X share the same copula, it is sufficient to transport the margins of X : if Q_1 and Q_2 transport W_1 into X_1 and W_2 into X_2 respectively (we recall that W_1 and W_2 are uniform such that we naturally choose Q_1 and Q_2 as the univariate quantiles of X_1 and X_2) then the function defined for $u, v \in [0;1]$ by

$$Q(u, v) = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} \circ Q_C(u, v) \tag{2.33}$$

transports U into X. To sum up, if we manage to transport a copula, we can easily transport all distributions sharing this copula.
We can link the copula function to the potential of Proposition 2.10 :

Lemma 2.6. *Let C be a copula and $Q_C = \nabla\phi_C$ the monotone transport between the uniform distribution function and the distribution whose cdf is C. Then for all $u, v \in [0;1]$:*

$$C(u,v) = vol\left((\nabla\phi_C)^{-1}([0;u]\times[0;v])\right). \tag{2.34}$$

Proof. Let $W = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$ be the distribution of cdf C. By definition, we have $W \overset{d}{=} \nabla\phi_C(U)$ (U is uniform on $[0;1]^2$).
Then, if $u, v \in [0;1]$

$$C(u,v) = \mathbb{P}[W_1 \le u, W_2 \le v] = \mathbb{P}[\partial_1\phi_C(U) \le u, \partial_2\phi_C(U) \le v] = vol\left((\nabla\phi_C)^{-1}([0;u]\times[0;v])\right).$$

□

Example 2.17. *(Independent Copula)*
The case of the independent copula $\Pi(u,v) = uv$ is straightforward. In that case, the potential is $\phi_\Pi(u,v) = \frac{u^2}{2} + \frac{v^2}{2}$ which gives :

$$Q_\Pi(u,v) = \nabla\phi_\Pi(u,v) = \begin{pmatrix} u \\ v \end{pmatrix}. \tag{2.35}$$

As (U,V) are uniform independent, $Q_\Pi(u,v)$ have independent margins and its copula is Π. ϕ_Π is then the associated potential for the independent copula.
The L-moments of the copula's distribution are then for $j,k > 0$:

$$\lambda_{jk} = \int_{[0;1]^2} \begin{pmatrix} u \\ v \end{pmatrix} L_j(u)L_k(v)dudv = \begin{pmatrix} \mathbb{1}_{k=1}\lambda_j(U) \\ \mathbb{1}_{j=1}\lambda_k(U) \end{pmatrix}$$

where U represents a uniform distribution on [0;1] i.e.

$$\lambda_{11} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \ \lambda_{12} = \begin{pmatrix} 0 \\ \frac{1}{6} \end{pmatrix}, \ \lambda_{21} = \begin{pmatrix} \frac{1}{6} \\ 0 \end{pmatrix}, \ \lambda_{jk} = 0 \text{ otherwise.}$$

Example 2.18. *(Max-Copula, continued)*
The copula-max $M(u,v) = \min(u,v)$ was treated in Example 2.16. The associated transport is :

$$Q_M(u,v) = \frac{1}{2}\begin{pmatrix} u+v \\ u+v \end{pmatrix}. \tag{2.36}$$

The L-moments of the copula's distribution are then for $j,k > 0$:

$$\lambda_{jk} = \frac{1}{2}\int_{[0;1]^2} \begin{pmatrix} u+v \\ u+v \end{pmatrix} L_j(u)L_k(v)dudv = \frac{1}{2}\begin{pmatrix} \mathbb{1}_{k=1}\lambda_j(U) + \mathbb{1}_{j=1}\lambda_k(U) \\ \mathbb{1}_{k=1}\lambda_j(U) + \mathbb{1}_{j=1}\lambda_k(U) \end{pmatrix}$$

where U represents a uniform distribution on $[0;1]$ i.e.

$$\lambda_{11} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \ \lambda_{12} = \begin{pmatrix} \frac{1}{12} \\ \frac{1}{12} \end{pmatrix}, \ \lambda_{21} = \begin{pmatrix} \frac{1}{12} \\ \frac{1}{12} \end{pmatrix}, \ \lambda_{jk} = 0 \text{ otherwise.}$$

Example 2.19. *(Min-Copula)*
The case of the copula-min $W(u,v) = \max(u+v-1, 0)$ *can similarly be solved. Let us define the potential* $\phi_W(u,v) = \frac{1}{4}(u+1-v)^2$*, then for* $u, v \in [0;1]$ *:*

$$Q_W(u,v) = \nabla\phi_W(u,v) = \frac{1}{2}\begin{pmatrix} u+1-v \\ v+1-u \end{pmatrix} \tag{2.37}$$

If U *and* V *are uniform and independent* $Q_W(U,V)$ *has uniform margins that are anti-comonotone i.e. the copula of* $Q_W(U,V)$ *is* W*. The L-moments of the copula are then for* $j,k > 0$ *:*

$$\lambda_{jk} = \frac{1}{2}\int_{[0;1]^2}\begin{pmatrix} u+1-v \\ v+1-u \end{pmatrix} L_j(u)L_k(v)dudv = \frac{1}{2}\begin{pmatrix} \mathbb{1}_{k=1}\lambda_j(U) + \mathbb{1}_{j=1}(-1)^k\lambda_k(U) \\ -\mathbb{1}_{k=1}(-1)^j\lambda_j(U) - \mathbb{1}_{j=1}\lambda_k(U) \end{pmatrix}$$

where U *represents a uniform distribution on [0 ;1] i.e.*

$$\lambda_{11} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix},\ \lambda_{12} = \begin{pmatrix} \frac{1}{12} \\ -\frac{1}{12} \end{pmatrix},\ \lambda_{21} = \begin{pmatrix} -\frac{1}{12} \\ \frac{1}{12} \end{pmatrix},\ \lambda_{jk} = 0 \text{ otherwise.}$$

For the sake of simplicity, we have presented copulas in the bivariate setting but multivariate generalizations of copulas and of Sklar's theorem exist. It is then straightforward to adapt the above transport for multivariate random vectors.

Remark 2.25. *Even if it is difficult to find the explicit formulation of the monotone transport for classical parametric family of copulas (such as Gumbel or Clayton copula), we can define a copula from its potential.*

2.4.3 Monotone transport from the standard Gaussian distribution

The major drawback of the uniform law on $[0;1]^d$ is its non-invariance by rotation which is a desirable property in order to more easily compute the monotone transports. For example, the multivariate standard Gaussian distribution appears as a better source measure but any other distribution could also be considered.
We propose an alternative transport leading to the following L-moments :

$$\lambda_\alpha = \int_{[0;1]^d} T_0 \circ Q_{\mathcal{N}}(t_1,...,t_d)L_\alpha(t_1,...,t_d)dt_1...dt_d \tag{2.38}$$

where $Q_{\mathcal{N}}$ is the transport of the multivariate standard distribution $\mathcal{N}(0, I_d)$ into the uniform one defined by

$$Q_{\mathcal{N}}(t_1,..,t_d) = \begin{pmatrix} \mathcal{N}^{-1}(t_1) \\ \vdots \\ \mathcal{N}^{-1}(t_d) \end{pmatrix}$$

and T_0 the transport of the considered distribution into the multivariate standard distribution :

$$([0;1], du) \overset{Q_{\mathcal{N}}}{\to} (\mathbb{R}^d, d\mathcal{N}) \overset{T_0}{\to} (\mathbb{R}^d, d\nu) \tag{2.39}$$

In [96], Sei used the transport from the standard Gaussian in order to define distributions through a convex potential φ (actually, he proposed to take the dual potential in the sense of Legendre duality). A useful property of this transport is given by the following lemma

Lemma 2.7. *Let $A \in O_d(\mathbb{R})$ be the space of orthogonal matrices (i.e. $AA^T = A^TA = I_d$), $m \in \mathbb{R}^d$, $a \in \mathbb{R}^*$. Let us denote by ϕ the potential linked to the random variable such that $\nabla\phi(N) \overset{d}{=} X$ with $N \in \mathbb{R}^d$ a standard Gaussian random vector.*
Then the respective potentials related to the vectors AX, aX and $X+m$ are $\phi(Ax)$, $a\phi(x)$ and $\phi(x)+m.x$.

Proof. Let $\psi_A(x) = \phi(Ax)$, then ψ is convex and $\nabla\psi_A(x) = A\phi(Ax)$. Furthermore, as A is an orthogonal matrix, $AN \overset{d}{=} N$ which implies $\nabla\psi_A(N) \overset{d}{=} AX$.
In the same way, if $\psi_a(x) = a\phi(x)$ and $\psi_m(x) = \phi(x) + m$, we have

$$\begin{aligned} \nabla\psi_a(N) &= a\nabla\phi(N) \overset{d}{=} aX \\ \nabla\psi_m(N) &= \nabla\phi(N) + m \overset{d}{=} X + m. \end{aligned}$$

□

Unfortunately, the generalization to all affine transformations is not easy. This is why it is often more convenient to define distributions through their potential function as in Sei's article.

Example 2.20. *(L-moments of multivariate Gaussian)*
Let us consider $m \in \mathbb{R}^d$, a positive matrix A and the quadratic potential :

$$\varphi(x) = m.x + \frac{1}{2}x^TAx \qquad \text{for } x \in \mathbb{R}^d. \tag{2.40}$$

The transport associated to this potential is :

$$T_0(x) = \nabla\varphi(x) = m + Ax \qquad \text{for } x \in \mathbb{R}^d. \tag{2.41}$$

Furthermore, $T_0(\mathcal{N}_d(0, I_d)) \overset{d}{=} \mathcal{N}_d(m, A^TA)$. The L-moments of a multivariate Gaussian of mean m and covariance A^TA are :

$$\begin{aligned} \lambda_\alpha &= \int_{[0;1]^d} [m + A\mathcal{N}_d(t_1, ..., t_d)]\, L_\alpha(t_1, ..., t_d)dt_1...dt_d \\ &= \mathbb{1}_{\alpha=(1,...,1)}m + \mathbb{1}_{\alpha\neq(1,...,1)}A\lambda_\alpha(\mathcal{N}_d(0, I_d)) \end{aligned}$$

with the notation $\lambda_\alpha(\mathcal{N}_d(0, I_d))$ denoting the α-th L-moments of the standard multivariate Gaussian, which is easy to compute since it is a random vector with independent components (see example 2.15).
In particular, the L-moment matrix of degree 2 :

$$\Lambda_2 = (\lambda_{2,1...,1} \dots \lambda_{1,...,1,2}) = A \begin{pmatrix} \frac{1}{\sqrt{\pi}} & 0 & \dots \\ 0 & \ddots & 0 \\ \dots & 0 & \frac{1}{\sqrt{\pi}} \end{pmatrix}. \tag{2.42}$$

The matrix of L-moments ratio of degree 2 is then

$$\tau_2 = (\tau_{2,1,...,1} \dots \tau_{1,...,1,2}) = \begin{pmatrix} \frac{a_{11}}{\left(\sum_{i=1}^d a_{i1}^2\right)^{1/2}} & \dots & \frac{a_{1d}}{\left(\sum_{i=1}^d a_{id}^2\right)^{1/2}} \\ \vdots & \ddots & \vdots \\ \frac{a_{d1}}{\left(\sum_{i=1}^d a_{i1}^2\right)^{1/2}} & \dots & \frac{a_{dd}}{\left(\sum_{i=1}^d a_{id}^2\right)^{1/2}} \end{pmatrix} \tag{2.43}$$

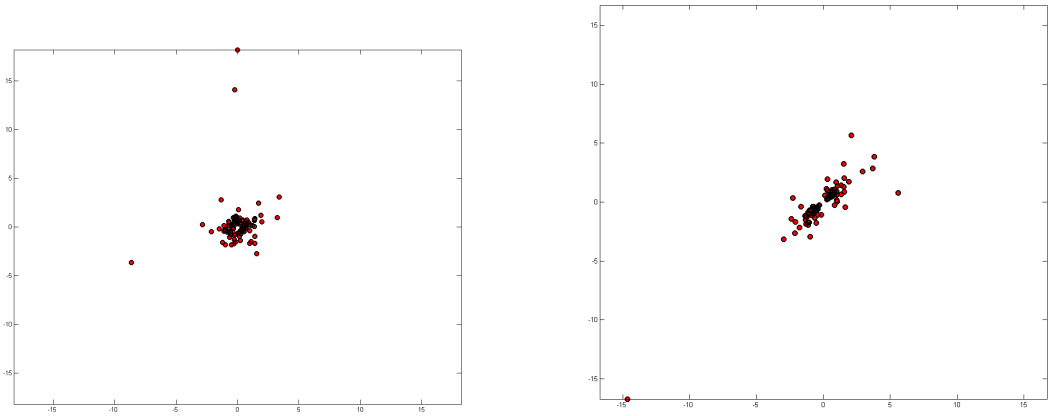


FIGURE 2.4 – Samples from the distribution induced by $T_0(X) = u'(x^T Ax)Ax$ with $u(x) = -\log(x)$ and $A = I_d$ (left) or $A = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$ (right)

with

$$A = \begin{pmatrix} a_{11} & \dots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{d1} & \dots & a_{dd} \end{pmatrix}$$

Example 2.21. *(Spherical and nearly-elliptical distributions)*
We now present a generalization of the previous example close to the elliptical family. Let $u : \mathbb{R} \to \mathbb{R}$ be a derivable strictly convex function,$m \in \mathbb{R}$, A be a positive matrix and define the potential :

$$\varphi(x) = m.x + \frac{1}{2}u(x^T Ax) \qquad \text{for } x \in \mathbb{R}^d. \tag{2.44}$$

The associated transport is given by :

$$T_0(x) = m + u'(x^T Ax)Ax. \tag{2.45}$$

If the integral is well defined, the L-moments of this distribution are then

$$\lambda_\alpha = \mathbb{1}_{\alpha=(1,\dots,1)}m + A\int_{\mathbb{R}^d} u'(x^T Ax)xL_\alpha(\mathcal{N}(x))d\mathcal{N}(x). \tag{2.46}$$

If we take $A = I_d$ and write $u'(x) = \frac{v(x)}{x^{1/2}}$, then $T_0(X) = m + v(X^T X)\frac{X}{(X^T X)^{1/2}}$ where X is a standard Gaussian random variable which is the characterization of a spherical distribution according to [32].

Example 2.22. *(Linear combinations of independent variables)*
Let $(e_1, ..., e_d)$ be an orthonormal basis of $\mathbb{R}^d$ and $(b_1, ..., b_d)$ the canonical basis. We consider the potential defined by :

$$\varphi(x) = \sum_{i=1}^{d} \sigma_i \varphi_i(x^T e_i) \tag{2.47}$$

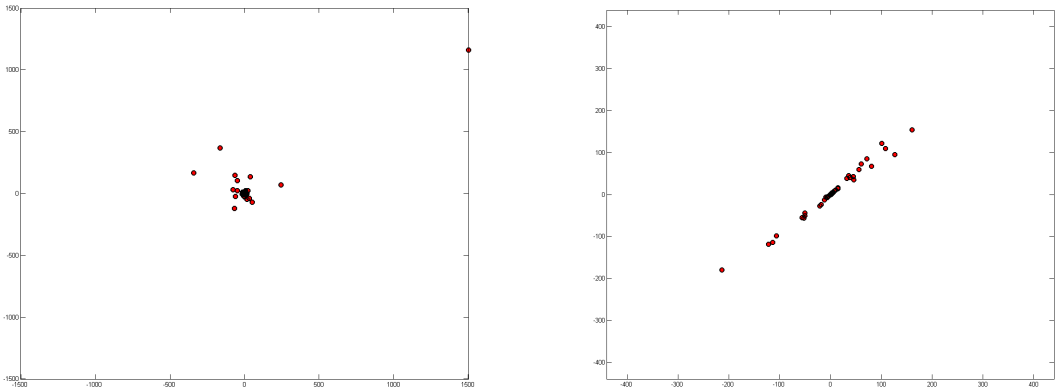


FIGURE 2.5 – Samples from the distribution induced by $T_0(X) = u'(x^T Ax)Ax$ with $u(x) = \frac{1}{3}x^3$ and $A = I_d$ (left) or $A = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$ (right)

with each function φ_i derivable and convex and $\sigma_i > 0$. Then

$$\nabla\varphi(x) = \sum_{i=1}^{d} \sigma_i \varphi_i'(x^T e_i) e_i \tag{2.48}$$

Then, if we denote by $P = \sum_{i=1}^d e_i b_i^T$ and $D = \sum_{i=1}^d \sigma_i b_i b_i^T$, this potential generates the random vector

$$Y \overset{d}{=} P^T D \begin{pmatrix} \varphi_1'(X^T e_1) \\ \vdots \\ \varphi_d'(X^T e_d) \end{pmatrix}. \tag{2.49}$$

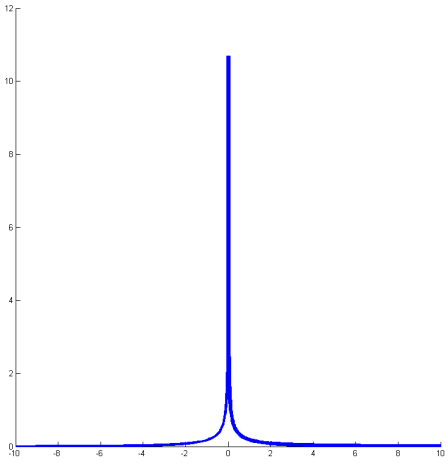
Let us note that P is orthogonal i.e. $PP^T = P^T P = I_d$ and D is diagonal.
As e_1,...,e_d is an orthonormal family, $X^T e_1, ..., X^T e_d$ are independent Gaussian random variables. Then if we write the increasing functions $\varphi_i'(x) = Q_i(\mathcal{N}_1(x))$ with Q_i the quantile of a random variable Z_i, then

$$Y \overset{d}{=} P^T \begin{pmatrix} \sigma_1 Z_1 \\ \vdots \\ \sigma_d Z_d \end{pmatrix} \tag{2.50}$$

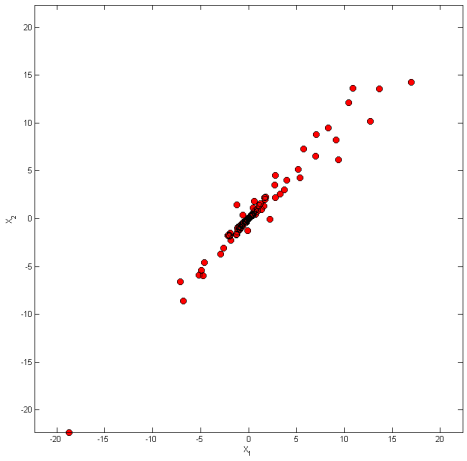
with Z_1,...,Z_d independent. The parameters σ_i are meant to represent a scale parameter for each Z_i but can be absorbed in the function φ_i'.
Figure 2.6 illustrates this model with for each i, $Z_i = \epsilon Z_i'$ where ϵ is a Rademacher random variable (i.e. discrete with probability $\frac{1}{2}$ on -1 and 1) and Z_i' is a Weibull random variable.

The L-moments of Y are then for $\alpha \in \mathbb{N}_*^d$:

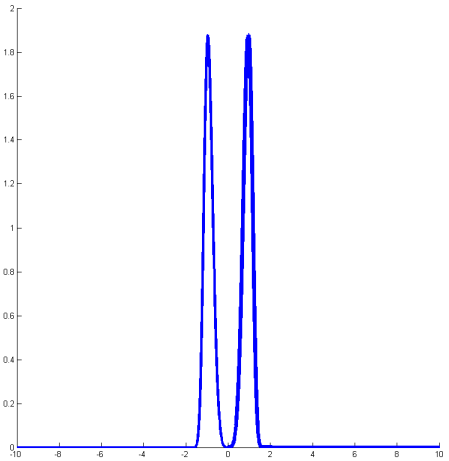
$$\lambda_\alpha = P^T D \begin{pmatrix} \int_{\mathbb{R}^d} \varphi_1'(\langle x, e_1\rangle) L_\alpha(\mathcal{N}_d(x)) d\mathcal{N}_d(x) \\ \vdots \\ \int_{\mathbb{R}^d} \varphi_d'(\langle x, e_d\rangle) L_\alpha(\mathcal{N}_d(x)) d\mathcal{N}_d(x)) \end{pmatrix}$$



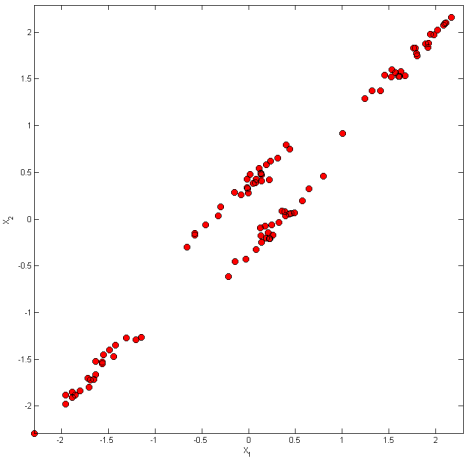
(a) Symmetrized Weibull density for Z_1 and Z_2 (shape parameter 0.5 and scale parameter 1)



(b) Corresponding samples



(c) Symmetrized Weibull density for Z_1 and Z_2 (shape parameter 2 and scale parameter 1)



(d) Corresponding samples

FIGURE 2.6 – Samples from the distribution given by equation (2.50) with $P = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, $\sigma_1 = 1.8$ and $\sigma_2 = 0.2$ (right) for different parameters for the Weibull distribution

2.5 Rosenblatt transport and L-moments

2.5.1 General multivariate case

In a paper dated of 1952 [93], Rosenblatt defined a transformation for the random variable $X = (X_1, ..., X_d)$ with an absolutely continuous distribution. This transformation denoted by T is now known as Rosenblatt transport (sometimes named Knothe's transport) and is explicitly given by the successive conditional distributions of $X_k|X_1 = x_1, ..., X_{k-1} = x_{k-1}$:

$$T(x_1, ..., x_d) = \begin{pmatrix} F_{X_1}(x_1) \\ F_{X_2|X_1}(x_2|x_1), \\ \vdots \\ F_{X_d|X_1,...,X_{d-1}}(x_d|x_1, ..., x_{d-1}) \end{pmatrix} \tag{2.51}$$

Rosenblatt showed that T transports the random variable X into the uniform law on $[0;1]^d$. However, T is not uniquely defined because there are $d!$ transports T corresponding to the $d!$ ways in which one can number the coordinates $X_1, ..., X_d$.
In the following, we soften the absolute continuity assumption in order to transport the uniform measure on $[0;1]^d$ μ onto an arbitrary measure ν. In that version, the Rosenblatt transport is based on the disintegration theorem (given without proof) which is a consequence of the Radon-Nikodym theorem (see for example [4]) :

Theorem 2.6. *Let E_1 and E_2 be two separable metric spaces equipped with their Borel σ-algebras, B_{E_1} and B_{E_2}. Let γ be a Borel probability measure on $E_1 \times E_2$ and $\gamma_1 = \pi_{E_1}\gamma$ be its first marginal; then there exists a family of probability measures on E_2, $(\gamma_2^{x_1})_{x_1 \in E_1}$ measurable in the sense that $x_1 \mapsto \gamma_2^{x_1}(A_2)$ is μ-measurable for every $A_2 \in B_{E_2}$ and such that $\gamma = \gamma_1 \otimes \gamma_2^{x_1}$ i.e. :*

$$\gamma(A_1 \times A_2) = \int_{A_1} \gamma_2^{x_1}(A_2) d\gamma_1(x_1) \tag{2.52}$$

for every $A_1 \in B_{E_1}$ and $A_2 \in B_{E_2}$.

We can sum up the previous theorem by stating the existence of measures $\gamma_2^{x_1}$ such that

$$\gamma = \gamma_1 \otimes \gamma_2^{x_1}. \tag{2.53}$$

$\gamma_2^{x_1}$ correspond to the notion of conditional distribution of the second marginal of γ knowing the first marginal is equal to x_1. The disintegration can be a way to define conditioning according to Chang and Pollard [38]. If γ is absolutely continuous and we have denoted its density by p, the disintegrated measures γ_1 and $\gamma_2^{x_1}$ have respective densities :

$$p_1(x_1) := \int p(x_1, x_2) dx_2 \quad \text{and} \quad p_2^{x_1}(x_2) := \frac{p(x_1, x_2)}{p_1(x_1)}.$$

The Rosenblatt transport refer to the concatenation of univariate transports of disintegrated measures from ν. More precisely in the case of the quantile, we recall that ν is a probability measure defined on the Borelian of $\mathbb{R}^d$. Let denote ν_1, $\nu_2^{x_1}$, ..., $\nu_d^{x_1,...,x_{d-1}}$ the disintegration of ν and F_1, $F_2^{x_1}$,..., $F_d^{x_1,...,x_{d-1}}$ the corresponding cdf. Then the Rosenblatt quantile is defined by

$$Q(t_1, t_2, ..., t_d) := \begin{pmatrix} Q_1(t_1) \\ Q_2(t_1, t_2) \\ \vdots \\ Q_d(t_1, ..., t_d) \end{pmatrix} = \begin{pmatrix} F_1^{-1}(t_1) \\ (F_2^{Q_1(t_1)})^{-1}(t_2) \\ \vdots \\ (F_d^{Q_1(t_1),...,Q_{d-1}(t_1,...,t_{d-1})})^{-1}(t_d) \end{pmatrix}. \tag{2.54}$$

This construction transports the uniform distribution on $[0;1]^d$ into the distribution of the random vector X and can be defined even if the distribution of X is not absolutely continuous.

Proposition 2.13. *If U is the uniform distribution on $[0;1]^d$ and Q is a Rosenblatt quantile, then $Q(U) \overset{d}{=} X$.*

Proof. We will prove the statement in the bivariate case for $Q = Q_{12}$ with

$$Q_{12}(t_1, t_2) = \begin{pmatrix} Q_1(t_1) \\ Q_2(t_1, t_2) \end{pmatrix}$$

The generalization to the multivariate case is very similar.
Let a be a function dF-measurable, then :

$$\begin{aligned}
\int_{[0;1]^2} a(Q_{12}(u))du &= \int_{[0;1]^2} a(Q_1(u), (F_2^{Q_1(u)})^{-1}(v))dudv \\
&= \int_0^1 \int_{\mathbb{R}} a(Q_1(u), y)dF_2^{Q_1(u)}(y)du \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} a(x, y)dF_2^x(y)dF_1(x) \\
&= \int_{\mathbb{R}^2} a(x, y)dF(x, y).
\end{aligned}$$

The second and third equalities hold because the successive quantiles are one-dimensional transports. □

Remark 2.26. *Carlier et al. [34] showed that the Rosenblatt transport can be viewed as a limit of optimal transports. Indeed, they showed that if we consider the cost depending on a parameter θ :*

$$c_\theta(x, y) = \frac{1}{2} \sum_{k=1}^{d} \theta^{k-1} |x_k - y_k|^2$$

then the mapping T_θ solving the optimal transport with such a cost converges in L^2 to the Rosenblatt transport given by equation (2.51) as θ goes to 0. We see once again that the Rosenblatt transport depends on the numbering order of the coordinates $x_1, .., x_d$ because c_θ is not symmetric with respect to the coordinates of x and y.

2.5.2 The case of bivariate L-moments of the form λ_{1r} and λ_{r1}

We now consider a bivariate vector $X = (X_1, X_2)$. The two possible Rosenblatt quantiles are given by the successive conditional quantiles

$$Q_{12}(t_1, t_2) = \begin{pmatrix} Q_{X_1}(t_1) \\ Q_{X_2|X_1=Q_{X_1}(t_1)}(t_2) \end{pmatrix}$$

or

$$Q_{21}(t_1, t_2) = \begin{pmatrix} Q_{X_1|X_2=Q_{X_2}(t_2)}(t_1) \\ Q_{X_2}(t_2) \end{pmatrix}$$

where Q_{X_1}, Q_{X_2} are the marginal quantiles of X_1 and X_2 and $Q_{X_2|X_1}, Q_{X_1|X_2}$ are the conditional quantiles.

The associated L-moments are then :

$$\begin{aligned} & \lambda_{\alpha}^{(12)} = \int_{[0;1]^2} Q_{12}(t_1,t_2)L_{\alpha}(t_1,t_2)dt_1dt_2 \\ \text{or} \quad & \lambda_{\alpha}^{(21)} = \int_{[0;1]^2} Q_{21}(t_1,t_2)L_{\alpha}(t_1,t_2)dt_1dt_2 \end{aligned}$$

Here, the multi-indices α are couples (r,s) for $r,s \geq 1$.
If we consider the pairs $(r,1)$ and $(1,s)$ and denote by $\lambda_r(X_i)$ the r-th univariate L-moment of X_i, we can express the corresponding L-moments :

$$\lambda_{r1}^{(12)} = \int_{[0;1]^2} Q_{12}(t_1,t_2)L_r(t_1)dt_1dt_2 = \begin{pmatrix} \lambda_r(X_1) \\ \mathbb{E}[L_r \circ F_1(X_1)\mathbb{E}[X_2|X_1]] \end{pmatrix} \tag{2.55}$$

and

$$\lambda_{1s}^{(21)} = \int_{[0;1]^2} Q_{21}(t_1,t_2)L_s(t_2)dt_1dt_2 = \begin{pmatrix} \mathbb{E}[L_s \circ F_2(X_2)\mathbb{E}[X_1|X_2]] \\ \lambda_s(X_2) \end{pmatrix}. \tag{2.56}$$

Serfling and Xiao [99] implicitly used this transformation for a bivariate vector to define multivariate L-moments. For a multivariate vector $X = (X_1, ..., X_d)$, they considered each pair $(X_i, X_j)_{1 \leq i,j \leq d}$ which avoids considering the $d!$ ways to build the Rosenblatt transport and allows a straightforward estimation through the concomitants of the samples as we will see in the next section.
They named r-th multivariate L-moments as the $d \times d$ matrix Λ_r

$$\Lambda_r = \begin{pmatrix} \Lambda_{r,11} & \Lambda_{r,12} & \dots & \Lambda_{r,1d} \\ \Lambda_{r,21} & \Lambda_{r,22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \Lambda_{r,d1} & \dots & \dots & \Lambda_{r,dd} \end{pmatrix}$$

defined so that each 2×2 submatrix is the concatenation of the above 2×1 vectors :

$$\begin{pmatrix} \Lambda_{r,ii} & \Lambda_{r,ij} \\ \Lambda_{r,ji} & \Lambda_{r,jj} \end{pmatrix} = \begin{pmatrix} \lambda_{1r}^{(ij)} & \lambda_{r1}^{(ji)} \end{pmatrix} \tag{2.57}$$

Example 2.23. *Unfortunately, these matrices are not sufficient for a total determination of a multivariate distribution. Let us present a copula that is an example of this assertion. Let $\theta \in [-1;1]$ and $C_\theta(u,v) = uv + \theta K_a(u)K_b(v)$ for $u,v \in [0;1]$ with $a,b \geq 3$. C is a copula because :*

- *$C(1,v) = v$ for all $v \in [0;1]$, $C(u,1) = u$ for all $u \in [0;1]$ and $C(u,0) = C(0,v) = 0$ for all $u,v \in [0;1]$*
- *if $u_1 \leq u_2$ and $v_1 \leq v_2$:*

$$\begin{aligned} & C(u_2,v_2) - C(u_1,v_2) - C(u_2,v_1) + C(u_1,v_1) \\ & \quad = (u_2-u_1)(v_2-v_1) + \theta(K_a(u_2) - K_a(u_1))(K_b(v_2) - K_b(v_1)) \\ & \quad \geq (1-\theta)(u_2-u_1)(v_2-v_1) \geq 0 \end{aligned}$$

because for all $a \geq 1$, K_a is 1-Lipschitzian.

Furthermore, if we consider the matrices defined by Serfling and Xiao :

$$\Lambda_{r,11} = \Lambda_{r,11} = \lambda_r(U([0;1]))$$

and

$$\Lambda_{r,12} = \mathbb{E}[L_r \circ F_1(X_1)\mathbb{E}[X_2|X_1]] = \int_{[0;1]^2} vL_r(u)dC_\theta(u,v) = \mathbb{1}_{r=1}\frac{1}{2}.$$

Similarly,

$$\Lambda_{r,21} = \int_{[0;1]^2} uL_r(v)dC_\theta(u,v) = \mathbb{1}_{r=1}\frac{1}{2}.$$

Hence, the whole family of cdf's $(C_\theta)_{\theta\in[-1;1]}$ *admits the same matrices* Λ_r*.*

Property of $\lambda_{r1}^{(12)}$ and $\lambda_{1r}^{(21)}$

We will present properties for $\lambda_{1r}^{(21)}$ that can be easily extended to $\lambda_{r1}^{(12)}$. Although these specific L-moments do not completely characterize any bivariate distribution, they share some desirable properties.

Proposition 2.14. *Let us recall that the L-moments ratios are defined by (see Definition 2.7)*

$$\tau_{1r}^{(21)} := \frac{\lambda_{1r}^{(21)}}{\lambda(X_2)} \in \mathbb{R}^2$$

Then, we have for $k = 1, 2$

$$|\langle \tau_{12}^{(21)}, b_k\rangle| \leq 1 \tag{2.58}$$

where (b_1, b_2) is the canonical basis of $\mathbb{R}^2$.

Proof. We apply Proposition 2.8 with $V \overset{d}{=} X_2$. □

Let us suppose $X_1, X_2, ..., X_r$ r bivariate random samples. If we order the samples along the second coordinate i.e. $X_{2,(1:r)} \leq X_{2,(2:r)} \leq ... \leq X_{2,(r:r)}$, the remaining first coordinate $X_{1,(i:r)}$, paired with each $X_{2,(i:r)}$, is named the concomitant of $X_{2,(i:r)}$ (see Yang [112] for a general study of concomitants). Furthermore, note

$$X_{(i:r)}^{(21)} = \begin{pmatrix} X_{1,(i:r)} \\ X_{2,(i:r)} \end{pmatrix}.$$

The superscript (21) refers to the choice of X_2 as sorting coordinate. We can then have an analogue characterization of the multivariate L-moment as a linear combination of expectations of concomitants.

Proposition 2.15. *The r-th L-moment may be represented as*

$$\lambda_{1r}^{(21)} = \frac{1}{r}\sum_{j=0}^{r-1}(-1)^j\binom{j}{r-1}\mathbb{E}[X_{(r-j:r)}^{(21)}] \tag{2.59}$$

Proof. Let $i \leq r$. We have $\mathbb{E}[X_{1,(i:r)}^{(21)}] = r\mathbb{E}[X_1|X_2 = X_{2,(i:r)}]$ i.e. by analogy with standard order statistics

$$\mathbb{E}[X_{1,(i:r)}^{(21)}] = r\binom{i-1}{r-1}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x_1(F_2(x_2))^{i-1}(1-F_2(x_2))^{r-i}dF(x_1,x_2)$$

We continue the analogy with the dimension 1 to reorganize the coefficients and conclude. □

This characterization allows us to use L-statistics and U-statistics representation especially in order to build unbiased estimators (see [99]).

2.6 Estimation of L-moments

Let $x_1, ..., x_n$ be independently drawn from a common random variable $X \in \mathbb{R}^d$ of measure ν. We will note $\nu_n = \sum_{i=1}^n \delta_{x_{(i)}}$ the empirical measure. The estimation of multivariate L-moments is built from an estimation of the quantile function, say Q_n. This section considers the estimation of Q, leading to explicit formulas for Q_n. The L-moments are estimated by plug-in, through

$$\hat{\lambda}_\alpha = \int_{[0;1]^d} Q_n(t) L_\alpha(t) dt \tag{2.60}$$

The simplest idea for the estimation of Q is to build the transport between the continuous uniform distribution on $\Omega = [0;1]^d$ and the discrete measure ν_n which is possible for the considered transports.

2.6.1 Estimation of the Rosenblatt transport

The estimation of this transport is attractive due to its simplicity and its similarity with the univariate case.
We suppose that the sampling distribution is **absolutely continuous** with respect to the Lebesgue measure. Then for two random samples X_i and X_j, $\mathbb{P}[X_i = X_j] = 0$.
If we denote by Q_n the empirical quantile built from the construction of the equation 2.54 for $\nu = \nu_n$, then $Q_n : [0;1]^d \to \{x_1, ..., x_n\}$ is defined with probability 1 for all $1 \leq i \leq n$ by :

$$Q_n(u_1, ..., u_d) = x^{(1)}_{(i:n)} \quad \text{for any } u_1 \in \left[\frac{i-1}{n}; \frac{i}{n}\right), u_2, ..., u_d \in [0;1]$$

where $x^{(1)}_{(1:n)}, x^{(1)}_{(2:n)}, \ldots, x^{(1)}_{(n:n)}$ denote the samples sorted by their first coordinate. Recall that we call the $(d-1)$ last components of $x^{(1)}_{(i:n)}$ the concomitants of its first component [112].

Remark 2.27. *If the sampling distribution is discrete for example, the expression of the quantile will be more complicated since the law of $X_2|X_1 = x_1$ is not reduced to a single point.*

Therefore, a natural version for the estimated L-moments associated to the Rosenblatt quantile could be :

$$\hat{\lambda}_{(i_1,...,i_d)} = \int_{[0;1]^d} Q_n(u) L_{(i_1,...,i_d)}(u) du = \sum_{i=1}^n w_i^{(i_1)} x^{(1)}_{(i:n)} \tag{2.61}$$

where $(i_1, ..., i_d) \in \mathbb{N}_*^d$ and

$$w_i^{(i_1)} = \int_{(i-1)/n}^{i/n} L_{i_1}(u_1) du_1 \mathbb{1}_{i_2=1} ... \mathbb{1}_{i_d=1}$$

are the weights of the estimators of the i_1-th univariate L-moments. Therefore, this estimator has an interest only for L-moments of the form $\lambda_{i_1,1,...,1}$. We will restrict ourselves to this case for L-moments associated to Rosenblatt quantiles. Serfling and Xiao [99] proposed a slightly different estimator which is the unbiased version of the above estimator :

$$\hat{\lambda}^{(u)}_{(i_1,1,...,1)} = \sum_{i=1}^n v_i^{(i_1)} x^{(1)}_{(i:n)} \tag{2.62}$$

with

$$v_i^{(i_1)} = \sum_{j=0}^{\min(i-1,i_1-1)} (-1)^{i_1-1-j} \binom{j}{i_1-1} \binom{j}{i_1-1+j} \binom{j}{n-1}^{-1} \binom{j}{i-1}.$$

Moreover, the consistency of both estimators holds for bivariate random vectors but in general fails to hold for vectors of dimension $d > 2$:

Theorem 2.7. *If we define for all $u \in [0;1]^d$:*

$$Q^{(1)}(u) = \begin{pmatrix} Q_{X_1}(u_1) \\ Q_{X_2|X_1=Q_{X_1}(u_1)}(u_2) \\ \vdots \\ Q_{X_d|X_1=Q_{X_1}(u_1)}(u_d) \end{pmatrix} \tag{2.63}$$

then

$$\hat{\lambda}_{(i_1,1,\dots,1)} \overset{a.s.}{\to} \int_{[0;1]^d} Q^{(1)}(u) L_{(i_1,1,\dots,1)}(u) du = \lambda_{(i_1 1,\dots,1)} \tag{2.64}$$

and

$$\hat{\lambda}^{(u)}_{(i_1,1,\dots,1)} \overset{a.s.}{\to} \int_{[0;1]^d} Q^{(1)}(u) L_{(i_1,1,\dots,1)}(u) du = \lambda_{(i_1 1,\dots,1)}. \tag{2.65}$$

Proof. The convergence of the first coordinate of $\hat{\lambda}_{(i_1,\dots,i_d)}$ or $\hat{\lambda}^{(u)}_{(i_1,\dots,i_d)}$ directly comes from the univariate L-moments convergence results [64]. The $(d-1)$ remaining coordinate converge as an application of the theorem of convergence for the linear combinations of concomitants [112]. □

Remark 2.28. *An other idea for the estimation of the Rosenblatt L-moments is to consider the Rosenblatt construction of the quantile with a smoothed version of the empirical distribution. If this smoothed version (for example a kernel version) is absolutely continuous with respect to the Lebesgue measure, then consistency would hold.*

2.6.2 Estimation of a monotone transport

We will show here that the monotone transport from any absolutely continuous distribution onto a discrete one is the gradient of a piecewise linear function. We present the construction of the monotone transport of an absolutely continuous measure μ defined on $\mathbb{R}^d$ onto $\nu_n = \sum_{i=1}^n \delta_{x_i}$. Here, μ will typically be either the standard Gaussian measure on $\mathbb{R}^d$ or the uniform measure on $[0;1]^d$. We will denote by Ω the support of μ.

Power diagrams

Here, we briefly present power diagrams, a tool generalizing Voronoi diagrams and coming from computational geometry, which is useful for the representation of the discrete optimal transport.

Definition 2.11. *Let $x_1, ..., x_n \in \mathbb{R}^d$ and their associated weights $w_1, ..., w_n \in \mathbb{R}$. The power diagram of $(x_1, w_1), ..., (x_n, w_n)$ is the subdivision of Ω into n polyhedra given by :*

$$\Omega = \bigcup_{1\le i\le n} PD_i = \bigcup_{1\le i\le n} \left\{ u \in \Omega \text{ s.t. } \|u - x_i\|^2 + w_i \le \|u - x_j\|^2 + w_j \quad \forall j \ne i \right\} \tag{2.66}$$

Remark 2.29. *If the weights are all zero and $x_1, ..., x_n \in \Omega$, then the power diagram is the Voronoi diagram.*

Convex piecewise linear functions are strongly related to power diagrams through their gradient. Indeed, let $\phi_h : \Omega \to \mathbb{R}$ be a piecewise linear function. Assume that ϕ_h is parametrized by $h = \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix} \in \mathbb{R}^n$. Define then ϕ_h explicitly through :

$$\text{for any } u \in \Omega, \quad \phi_h(u) = \max_{1 \leq i \leq n} \{u.x_i + h_i\}. \tag{2.67}$$

Let $(W_i(h))_{1 \leq i \leq n}$ be the polyhedron partition of Ω defined by

$$W_i(h) = \{u \in \Omega \text{ s.t. } \nabla\phi_h(u) = x_i\}.$$

This subdivision is often called the natural subdivision associated to the piecewise linear function ϕ_h. Then, we have the following lemma :

Lemma 2.8. *The power diagram associated to $(x_1, w_1), ..., (x_n, w_n)$ is the polyhedron partition $\cup_{1 \leq i \leq n} W_i(h)$ if $h_i = -\frac{\|x_i\|^2 + w_i}{2}$.*

Proof. The proof is straightforward since $\nabla\phi_h(u) = x_i$ iff $u.x_i + h_i \geq u.x_j + h_j$ for all j which is equivalent to $\|u - x_i\|^2 + 2h_i - \|x_i\|^2 \geq \|u - x_j\|^2 + 2h_j - \|x_j\|^2$. □

Discrete monotone transport

We will present a variational approach initially proposed by Aurenhammer [8] for the quadratic optimal transportation problem between a probability measure μ defined on Ω and the empirical distribution of a sample $x_1, ..., x_n$ which is denoted by ν_n.
Let $\phi_h : \Omega \to \mathbb{R}$ be the piecewise linear function defined by Equation (2.67).

Theorem 2.8. *Let us suppose that $x_1, ..., x_n$ are distinct points of $\mathbb{R}^d$. Let Ω be a convex domain of $\mathbb{R}^d$ such that $vol(\Omega) > 0$ and μ an absolutely continuous probability measure with finite expectation.*
Then $\nabla\phi_h$ is piecewise constant and is a monotone transport of μ into ν_n with a particular $h = h^*$, unique up to a constant $(b, ..., b)$, which is the minimizer of an energy function E

$$h^* = \arg\min_{h \in \mathbb{R}^n} E(h) = \arg\min_{h \in \mathbb{R}^n} \int_\Omega \phi_h(u) d\mu - \frac{1}{n} \sum_{i=1}^n h_i. \tag{2.68}$$

Furthermore, E is strictly convex on

$$H_0^{(n)} = \left\{ h \in \mathbb{R}^n \text{ s.t. for any } 1 \leq i \leq n \text{ , } W_i(h) \neq \emptyset \text{ and } \sum_{i=1}^n h_i = 0 \right\}.$$

Proof. The proof of Theorem 1.2 of Gu et al. can be extended to the case of a measure μ defined on an arbitrary convex set $\Omega \subset \mathbb{R}^d$. However, we do not prove that ∇E is a local diffeomorphism.
The proof is delayed to the Appendix. □

Remark 2.30. *The convexity of the domain Ω is needed in order to ensure that $H_0^{(n)}$ is non void.*

As exposed in the Appendix, the gradient is simply given by :

$$\nabla E(h) = \begin{pmatrix} \int_{W_1(h)} d\mu(x) - \frac{1}{n} \\ \vdots \\ \int_{W_n(h)} d\mu(x) - \frac{1}{n} \end{pmatrix}. \tag{2.69}$$

Moreover, for the expression of the Hessian of E, let us define the intersection faces for $1 \leq i, j \leq n$:

$$\begin{cases} F_{ij} = W_i(h) \cap W_j(h) \cap \Omega & \text{if the codimension of } F_{ij} \text{ is 1} \\ F_{ij} = \emptyset & \text{otherwise} \end{cases}$$

Then, if dA denote the area form on F_{ij}, the Hessian of E is given by

$$\begin{cases} \frac{\partial^2 E}{\partial h_i \partial h_j} = -\frac{1}{\|x_i - x_j\|} \int_{F_{ij}} dA & \text{if } i \neq j \\ \frac{\partial^2 E}{\partial h_i \partial h_i} = \sum_{1 \leq j \leq n, j \neq i} \frac{1}{\|x_i - x_k\|} \int_{F_{ij}} dA \end{cases}.$$

We can perform the computation of the solution h^* of the minimization problem by Newton's method.
If $\Omega = [0;1]^d$, in order to initialize this algorithm with $h^{(0)}(x_1, ..., x_n) \in H_0^{(n)}$, we consider the vector corresponding to the translation/scaling of the classical Voronoi cells into $[0;1]^d$ i.e. :

$$h^{(0)}(x_1, ..., x_n) = \frac{1}{4m_n} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n |x_i|^2 - |x_1|^2 \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n |x_i|^2 - |x_n|^2 \end{pmatrix}$$

with m_n the largest coordinate absolute value among the sample $x_1, .., x_n$.

Proposition 2.16. *If $\Omega = [0;1]^d$ and the x_i's are distinct :*

$$h^{(0)}(x_1, ..., x_n) \in H_0^{(n)}.$$

Proof. Let us define the hypercube englobing all the samples $X_n = [-m_n; m_n] \times \cdots \times [-m_n; m_n]$. Let $V_1, ..., V_n$ the Voronoi cells intersected with X_n. So with probability 1 :

$$V_i = \{x \in X_n \text{ s.t. } |x - x_i| \leq |x - x_j| \quad \forall j \neq i\} \neq \emptyset$$

It is clear that if $h_V = \begin{pmatrix} -\frac{1}{2}|x_1|^2 \\ \vdots \\ -\frac{1}{2}|x_n|^2 \end{pmatrix}$, we have

$$V_i = \{y \in X_n \text{ s.t. } \nabla \phi_{h_V}(y) = x_i\}$$

Let us note $u_c = \begin{pmatrix} 1/2 \\ \vdots \\ 1/2 \end{pmatrix}$. Then, if $h_\Omega = \frac{1}{2m_n} h_V - \begin{pmatrix} \langle x_1, u_c \rangle \\ \vdots \\ \langle x_n, u_c \rangle \end{pmatrix}$ and $u \in \Omega = [0;1]^d$:

$$\begin{aligned} \phi_{h_\Omega}(u) &= \max_{1 \leq i \leq n} \{\langle u, x_i \rangle + h_{\Omega,i}\} \\ &= \frac{1}{2m_n} \max_{1 \leq i \leq n} \{2m_n \langle (u - u_c), x_i \rangle + 2m_n (h_{\Omega,i} + \langle x_i, u_c \rangle)\} \\ &= \frac{1}{2m_n} \phi_{h_V}(2m_n (u - u_c)) \end{aligned}$$

So for any $1 \leq i \leq N$, $W_i(h_\Omega) = 2m_n (V_i - u_c) \neq \emptyset$.
We end this proof by taking as initialization vector $h^{(0)}(x_1, ..., x_n) = h_\Omega - \frac{1}{n} \sum_{i=1}^n h_{\Omega,i}$. □

Remark 2.31. *If $\Omega = \mathbb{R}^d$, this initialization is not an issue since it suffices to take the vector h corresponding to the Voronoi cells.*

Algorithm 1 Computation of the discrete optimal transport via Newton's method

Aim : To compute the discrete subdivision of Ω designing the optimal transport between ν_n and μ
Input : $h_0 \in H_0^{(n)}$, a descent step γ, a tolerance η
while $|\nabla E(h_t)| > \eta$
  $h_{t+1} = h_t - \gamma(\nabla^2 E(h_t))^{-1}\nabla E(h_t)$
  $t \leftarrow t+1$
end

In practice, the Hessian $\nabla^2 E$ is often hard to compute since it requires the calculation of the area of the facets of a power diagram. In our implementation, we prefer to use the simpler gradient descent in Algorithm 1 :

$$h_{t+1} = h_t - \gamma \nabla E(h_t).$$

Moreover, in order to compute the gradient of E for an arbitrary measure μ, we use a Monte-Carlo method.
However, since E is strictly convex only in $H_0^{(n)}$, Algorithm 1 may not converge to h^* especially when n is large. An improvement of this algorithm that would perform a gradient descent on the set $H_0^{(n)}$ is left as perspective.

Explicit expression for 2 samples

As an illustration, we can explicitly compute the monotone transport and some associated L-moments for 2 samples with a source distribution equal to the uniform on $[0;1]^d$ or to the standard normal $\mathcal{N}_d(0, I_d)$. Let $x_1, x_2 \in \mathbb{R}^d$ two samples coming from the same distribution.
Let us first begin with the standard normal distribution as source measure. As the potential ϕ_h of the previous section is defined up to an additive constant, we consider for $h \in \mathbb{R}^d$:

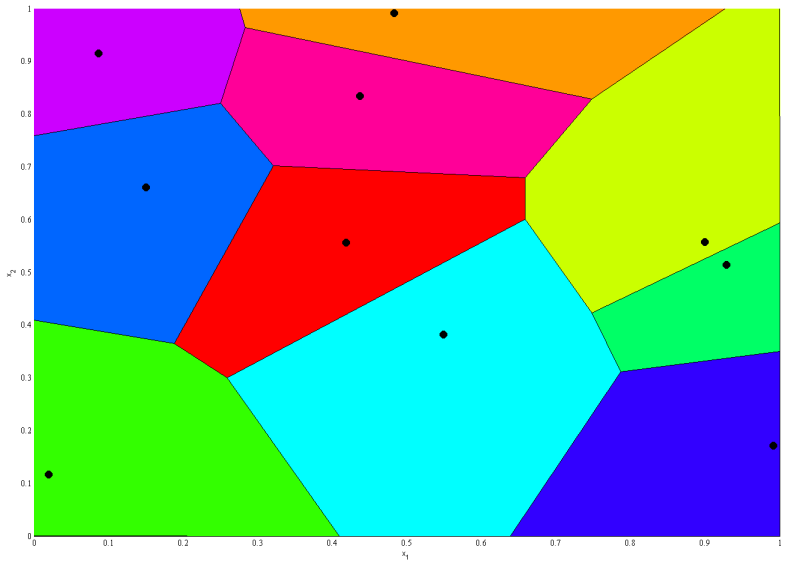
$$\phi_h(u) = \max(u.x_1, u.x_2 + h)$$

$\nabla\phi_h$ is the discrete optimal transport if $W_i = \{y \in \mathbb{R} \text{ s.t. } \nabla\phi_h(y) = x_i\}$ for $i = 1, 2$ have a measure equal to $1/2$ for the normal measure. By symmetry, we can assert that this property is attained for $h = 0$. The transport is then for $y \in \mathbb{R}$:

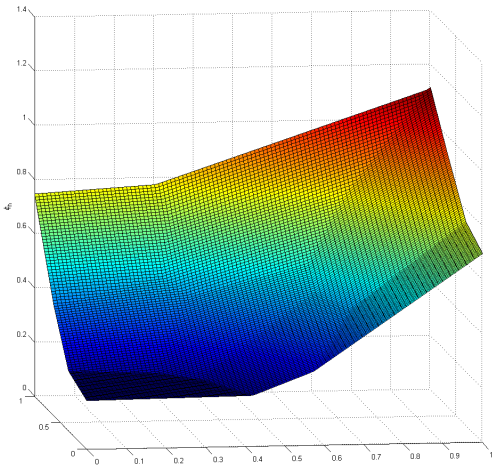
$$T_{\mathcal{N}}(y) = \nabla\phi_0(y) = \begin{cases} x_1 \text{ if } y.(x_1 - x_2) \geq 0 \\ x_2 \text{ if } y.(x_1 - x_2) \leq 0 \end{cases} \tag{2.70}$$

The L-moments of degree 2 associated with this transport are then (for the sake of simplicity, we compute only the L-moments related to the first coordinate) :

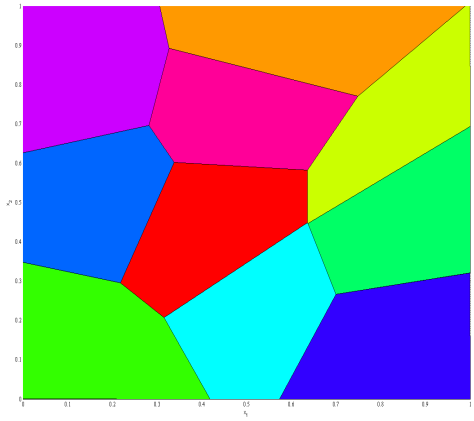
$$\begin{aligned} \lambda_{2,1,\dots,1}(x_1, x_2) &= \int_{\mathbb{R}^d} \nabla\phi_0(y) L_2(\mathcal{N}(y_1)) d\mathcal{N}_d(y) \\ &= (x_1 - x_2) \int_{y.(x_1 - x_2) \geq 0} L_2(\mathcal{N}(y_1)) d\mathcal{N}_d(y). \end{aligned}$$



(a) Voronoi cells of the sample

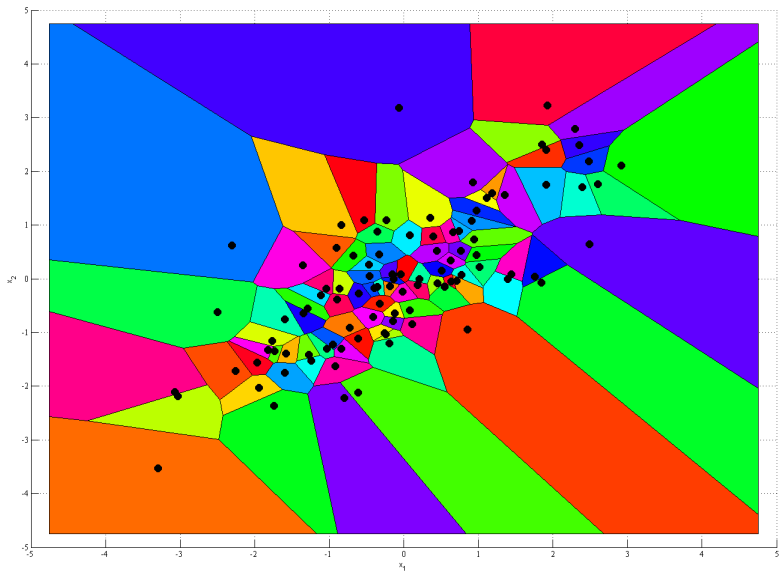


(b) Potential function of the optimal transport

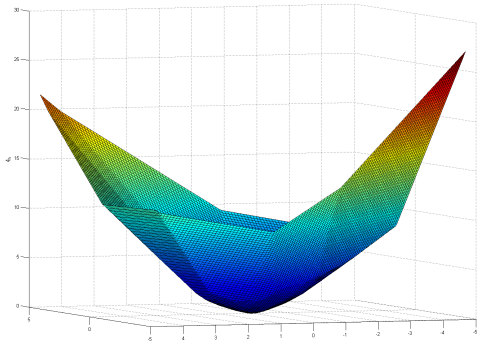


(c) Power diagram corresponding to the optimal transport (the transport maps each cell into one sample i.e. is piecewise linear)

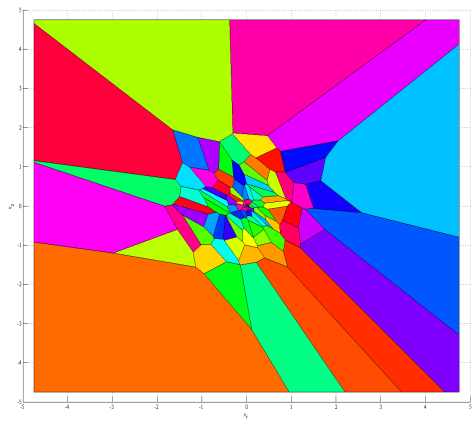
FIGURE 2.7 – Optimal transport of the discrete empirical distribution of a sample of size 10 into the uniform distribution on $[0;1]^2$



(a) Voronoi cells of the sample



(b) Potential function of the optimal transport



(c) Power diagram corresponding to the optimal transport (the transport maps each cell into one sample i.e. is piecewise linear)

Figure 2.8 – Discrete optimal transport for a sample of size 100 drawn from a Gaussian distribution with covariance $\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$ into the standard Gaussian

Let us denote by $(e_1, .., e_d)$ the canonical basis of $\mathbb{R}^d$. Then, if e_1 and $x_1 - x_2$ are collinear, then

$$\int_{y.(x_1-x_2)\geq 0} L_2(\mathcal{N}(y_1))d\mathcal{N}_d(y) = \pm 1/4$$

depending on the sign of $y.(x_1 - x_2)$. If it is not the case, we can build an orthonormal basis $(f_1 = e_1, f_2, ..., f_d)$ by completing the basis of the plan formed by e_1 and $x_1 - x_2$. Let us denote by U the rotation matrix transforming the canonical basis into the second one. We also define $a_1 = (x_1 - x_2).f_1 = (x_1 - x_2).e_1$ and $a_2 = (x_1 - x_2).f_2$. Then, the integral becomes

$$\begin{aligned}\lambda_{2,1,...,1}(x_1, x_2) &= (x_1 - x_2) \int_{a_1 z_1 + a_2 z_2 \geq 0} L_2(\mathcal{N}(z_1))d\mathcal{N}_d(z) \\ &= (x_1 - x_2) \int_{\mathbb{R}} L_2(\mathcal{N}(z_1))\mathcal{N}(-\frac{a_1}{|a_2|}z_1)d\mathcal{N}(z_1) \\ &= \frac{x_1 - x_2}{\pi} \arctan\left(\frac{a_1/|a_2|}{\sqrt{(a_1/|a_2|)^2 + 2}}\right)\end{aligned}$$

This last equality is obtained by deriving the function $t \mapsto \int L_2(\mathcal{N}(z_1))\mathcal{N}(tz_1)d\mathcal{N}(z_1)$ and is still valid for $a_2 = 0$ which corresponds to the case of collinearity of e_1 and $x_1 - x_2$.

In the following, we will note the central point of the unit square $u_c = \begin{pmatrix} 1/2 \\ \vdots \\ 1/2 \end{pmatrix}$. The same calculus can be performed when the source measure is uniform on the unit square. The transport is then by the same argument of symmetry for $u \in [0; 1]^d$:

$$T_{unif}(u) = \begin{cases} x_1 \text{ if } (u - u_c).(x_1 - x_2) \geq 0 \\ x_2 \text{ if } (u - u_c).(x_1 - x_2) \leq 0 \end{cases}.$$

Performing the same kind of change of coordinate with a translation of u_c, the L-moments of order $(r, 1, ..., 1)$ with $r \geq 2$ are :

$$\begin{aligned}\lambda_{r,1,...,1}(x_1, x_2) &= (x_1 - x_2) \int_{(u-u_c).(x_1-x_2)\geq 0} L_r(u_1)du \\ &= (x_1 - x_2) \int_{a_1 v_1 + a_2 v_2 \geq 0; |v_1|, |v_2| \leq 1/2} L_r(v_1 + 1/2)dv_1 dv_2 \\ &= \begin{cases} \text{sgn}(a_1)K_r(\frac{1}{2}) & \text{if } a_2 = 0 \\ \frac{a_1}{6|a_2|} & \text{if } \frac{|a_2|}{|a_1|} \geq 1 \\ \frac{a_2}{2|a_2|}(1 + \frac{a_1}{|a_1|})\left[K_r(\frac{1}{2}(1 + \left|\frac{a_2}{a_1}\right|)) - K_r(\frac{1}{2}(1 - \left|\frac{a_2}{a_1}\right|))\right] & \\ -\frac{a_1}{|a_2|}\left[J_r(\frac{1}{2}(1 + \left|\frac{a_2}{a_1}\right|)) - J_r(\frac{1}{2}(1 - \left|\frac{a_2}{a_1}\right|))\right] & \text{otherwise} \end{cases}\end{aligned}$$

where K_r and J_r are successive primitive functions of L_r.

Consistency of the optimal transport estimator

Let $X_1, ..., X_n$ be n independent copies of a vector X in $\mathbb{R}^d$ with distribution ν. Let μ denote a reference measure on a convex set $\Omega \subset \mathbb{R}^d$ so that μ gives no mass to small sets. Let ν_n be the empirical measure pertaining to the sample.
We define two transports, say T and T_n expressed as the gradient of two convex functions, say φ and φ_n, so that $T = \nabla\varphi$ and $T_n = \nabla\varphi_n$.

T and T_n respectively transport μ onto ν and ν_n. We do not assume that the hypothesis in Theorem 2.4 holds for μ. Hence neither T nor T_n can be defined as an optimal transport for a quadratic cost ; T and T_n are merely monotone transports.
This section is devoted to the statement of the convergence of T_n to T.

Definition 2.12. *A set $S \subset \mathbb{R}^d \times \mathbb{R}^d$ is said to be cyclically monotone if for any finite number of points $(x_i, y_i) \in S$, i=1...n*

$$\langle y_1, x_2 - x_1\rangle + \langle y_2, x_3 - x_2\rangle + \dots \langle y_n, x_1 - x_n\rangle \leq 0 \tag{2.71}$$

By extension, we say that a function f is cyclically monotone if all subsets of the form

$$S = \{(x_1, f(x_1)), ..., (x_n, f(x_n))\}$$

are cyclically monotone.

Before stating consistency results, let us first begin with a lemma.

Lemma 2.9. *Let K be the space defined by*

$$K = \{\nabla\varphi \in L_1(\Omega, \mathbb{R}^d, \mu) \ , \ \varphi \text{ convex } \mu\text{-a.e.}\} \tag{2.72}$$

Then K is a Hilbert space for the norm :

$$\|\nabla\varphi\|_1 = \int_\Omega \|\nabla\varphi(x)\| d\mu(x) \tag{2.73}$$

Proof. As $L_1(\Omega, \mathbb{R}^d, \mu)$ is a Hilbert space, it is sufficient to prove that K is closed in $L_1(\Omega, \mathbb{R}^d, \mu)$. Let $\nabla\varphi_n$ be a sequence in K convergent to $T \in L_1(\Omega, \mathbb{R}^d, \mu)$. $\nabla\varphi_n$ is cyclically monotone i.e. for all $m \in \mathbb{N}$ and $x_0, x_1, ..., x_m \in \Omega$:

$$(x_1 - x_0).\nabla\varphi_n(x_0) + (x_2 - x_1).\nabla\varphi_n(x_1) + \dots + (x_0 - x_m).\nabla\varphi_n(x_m) \leq 0$$

Let $n \to \infty$ then :

$$(x_1 - x_0).T(x_0) + (x_2 - x_1).T(x_1) + \dots + (x_0 - x_m).T(x_m) \leq 0$$

i.e. T is cyclically monotone. Furthermore, Theorem 24.8 of Rockafellar asserts that there exists a convex potential φ whose subgradient is cyclically monotone [91]. As μ gives no mass to small sets, φ is μ-almost everywhere differentiable (see [5]) and $\nabla\varphi = T \in K$. □

Lemma 2.10. *(Lemma 9 McCann [79])*
Let a sequence of probability measure on $\mathbb{R}^d \times \mathbb{R}^d$, denoted by γ_n, converge to γ in the sense that for any test function $h \in C^\infty(\mathbb{R}^d \times \mathbb{R}^d)$ with compact support

$$\int h(x, y) d\gamma_n(x, y) \to \int h(x, y) d\gamma_(x, y).$$

Let us call the marginals of γ the respective measures μ and ν defined on $\mathbb{R}^d$ such that for any Borel set M of $\mathbb{R}^d$

$$\begin{aligned} \mu(M) &= \gamma(M \times \mathbb{R}^d) \\ \nu(M) &= \gamma(\mathbb{R}^d \times M) \end{aligned}$$

Then

- *γ has a cyclically monotone support if γ_n does for each n*
- *if the marginals of γ_n, denoted by μ_n and ν_n converge in the sense given above to μ and ν, then μ and ν are the respective marginals of γ*

Proof. It is an application of McCann's Lemma 9 [79] □

Theorem 2.9. *If ν satisfies $\int \|x\| d\nu(x) < +\infty$, let T and T_n be the monotone transports (i.e. gradients of convex function) of μ into ν and ν_n. Then :*

$$\|T - T_n\|_1 = \int_\Omega \|T(x) - T_n(x)\| d\mu(x) \overset{a.s.}{\to} 0. \tag{2.74}$$

Proof. T is a gradient of a convex potential. We will consider the space

$$K = \{\nabla\varphi \in L_1(\Omega, \mathbb{R}^d, \mu) \ , \ \varphi \text{ convex } \mu\text{-a.e.}\}.$$

T_n and T respectively transport μ into ν_n and ν. By the strong law of large numbers :

$$\int_\Omega \|T_n(x)\| d\mu(x) = \int_{\mathbb{R}^d} \|y\| d\nu_n(y) \overset{a.s.}{\to} \int_{\mathbb{R}^d} \|y\| d\nu(y).$$

Let ω be a realization such that :

$$\int_\Omega \|T_n(\omega, x)\| d\mu(x) = \int_{\mathbb{R}^d} \|y\| d\nu_n(\omega, y) \to \int_{\mathbb{R}^d} \|y\| d\nu(\omega, y).$$

In the following, we will omit ω for the sake of simplicity of the notations.
We deduce from the convergence result of $\|T_n\|_1$ that T_n is bounded for n large enough. Hence, there exists $\nabla\psi \in K$ such that $T_m = \nabla\varphi_m \to \nabla\psi$ in K for a subsequence $\{m\}$. If we set $d\gamma_m(x, y) = \delta(y - \nabla\varphi_m(x))d\mu(x)$ and $d\gamma(x, y) = \delta(y - \nabla\psi(x))d\mu(x)$. Then for any function f with compact support,

$$\int f(x, y) d\gamma_m(x, y)(x) \to \int f(x, y) d\gamma(x, y)(x).$$

By the above Lemma 2.10, we have that γ have μ and ν as marginals, i.e. $\nabla\psi$ maps μ into ν. By the uniqueness of the gradient of the convex transport, $\nabla\psi = \nabla\varphi = T$ is the unique limit point of the sequence T_n in the Hilbert K. □

Let T and T_n be the transport of a reference measure μ_0 onto ν and ν_n and Q_0 the transport of the uniform measure on $[0; 1]^d$ onto this reference measure. Let us recall that we defined the quantiles of ν and ν_n by $Q = T \circ Q_0$ and $Q_n = T_n \circ Q_0$.

Theorem 2.10. *Let ν satisfy $\int \|x\| d\nu(x) < +\infty$. Then, we have for $\alpha \in \mathbb{N}_*^d$.*

$$\hat{\lambda}_\alpha = \int_\Omega Q_n(u) L_\alpha(u) du \overset{a.s.}{\to} \lambda_\alpha = \int_\Omega Q(u) L_\alpha(u) du \tag{2.75}$$

Proof. By using Theorem 2.9, we have :

$$\begin{aligned}
\|\hat{\lambda}_\alpha - \lambda_\alpha\| &= \left\| \int_{[0;1]^d} T_n(Q_0(t)) L_\alpha(t) dt - \int_{[0;1]^d} T(Q_0(t)) L_\alpha(t) dt \right\| \\
&\leq \left(\int_{[0;1]^d} \|T_n(Q_0(t)) - T(Q_0(t))\| dt \right) \sup_{t \in [0;1]^d} L_\alpha(t) \\
&\leq \int_{\mathbb{R}^d} \|T_n(x) - T(x)\| d\mu_0(x) \overset{a.s.}{\to} 0.
\end{aligned}$$

□

Remark 2.32. *The L-moment estimator presented above has a L-statistic representation. For* $\alpha \in \mathbb{N}_*^d$

$$\hat{\lambda}_\alpha = \int_{[0;1]^d} T_n(Q_0(t)) L_\alpha(t) dt = \sum_{i=1}^{n} \left(\int_{Q_0^{-1}(W_i(T_n))} L_{\alpha(t)} dt \right) x_{(i)}$$

where

$$W_i(T_n) = \left\{ x \in \mathbb{R}^d \ s.t. \ T_n(x) = x_i \right\}$$

2.7 Some extensions

2.7.1 Trimming

Semi-robust univariate trimmed L-moments

Elamir and Seheult [50] proposed a trimmed version of univariate L-moments. Let us recall some notations. If $X_1, ..., X_r$ are real-valued iid random variables, we note $X_{1:r} \leq X_{2:r} \leq \cdots \leq X_{r:r}$ the order statistics. The TL-moments of order r are defined for two trimming parameters t_1 and t_2 :

$$\lambda_r^{(t_1,t_2)} = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{k}{r-1} \mathbb{E}[X_{r-k+t_1:r+t_1+t_2}] \tag{2.76}$$

If $t_1 = t_2 = 0$, the trimmed L-moments reduce to standard L-moments. Intuitively, we do not consider the t_1 first lower samples and the t_2 higher.

Example 2.24. *Let us present some trimmed L-moments of low order :*

$$\begin{aligned}
\lambda_1^{(1,1)} &= \mathbb{E}[X_{2:3}] \\
\lambda_1^{(1,0)} &= \mathbb{E}[X_{1:2}] \\
\lambda_2^{(1,1)} &= \frac{1}{2}\mathbb{E}[X_{3:4} - X_{2:3}] \\
\lambda_3^{(1,1)} &= \frac{1}{3}\mathbb{E}[X_{4:5} - 2X_{3:5} + X_{2:5}] \\
\lambda_4^{(1,1)} &= \frac{1}{4}\mathbb{E}[X_{5:6} - 3X_{4:6} + 3X_{3:6} - X_{2:6}]
\end{aligned}$$

The expectations of the order statistics are written in function of the quantile of the common distribution of the X_i's through

$$\mathbb{E}[X_{i:r}] = \frac{r!}{(i-1)!(r-i)!} \int_0^1 Q(u) u^{i-1} (1-u)^{r-i} du$$

We may also give the alternative definition of the TL-moments as scalar product in $L_2([0;1])$:

$$\lambda_r^{(t_1,t_2)} = \int_0^1 Q(u) P_r^{(t_1,t_2)}(u) du$$

with

$$P_r^{(t_1,t_2)}(u) = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{k}{r-1} \frac{(r+t_1+t_2)!}{(r-k+t_1-1)!(t_2+k)!} u^{r-k+t_1-1} (1-u)^{t_2+k}$$

It is worth noting that $\lambda_r^{(t_1,t_2)}$ may exist even if the common distribution of X_i does not have a finite expectation. For example, the existence holds for Cauchy distribution of parameter $x_0 \in \mathbb{R}, \sigma > 0$ and $t_1, t_2 > 1$.
Also, some robustness of the sampled trimmed L-moments to outliers holds. Let $x_1, ..., x_n$ an iid sample, then the empirical TL-moments are defined by the U-statistics corresponding to Definition 2.76 taking into account all subsamples of size r, similarly to the empirical L-moments. Hence, we remark that the t_1 lower and the t_2 larger x_i's are not considered for the empirical TL-moments. Although this estimator is robust to these extreme points, the breakdown point of this sample version is 0 because it eliminates a fixed number of extreme values, and so the proportion of eliminated samples will be asymptotically zero. It can be interesting to suppress a fixed proportion of high value in order to reinforce the robustness of the tool.

An other approach for robust multivariate trimmed L-moments

We cannot directly adapt the univariate trimmed version of L-moments to the multivariate case. Indeed, the multivariate L-moments are not expressed as linear combinations of expectations of order statistics. We then propose the following definition for trimmed L-moments :

$$\lambda_\alpha^{(D)} = \int_D Q(t) L_\alpha(t) dt \tag{2.77}$$

with Q a transport of the uniform distribution in $[0;1]^d$ into the distribution of interest, L_α the multivariate Legendre polynomial of index α and D a domain included in $[0;1]^d$. For example, the most intuitive choice would be to consider $D = [t_1, 1-t_1] \times \cdots \times [t_d, 1-t_d]$ with $t_1, \ldots, t_d \in [0; \frac{1}{2}]$ representing **the proportion of extremal deleted samples**.

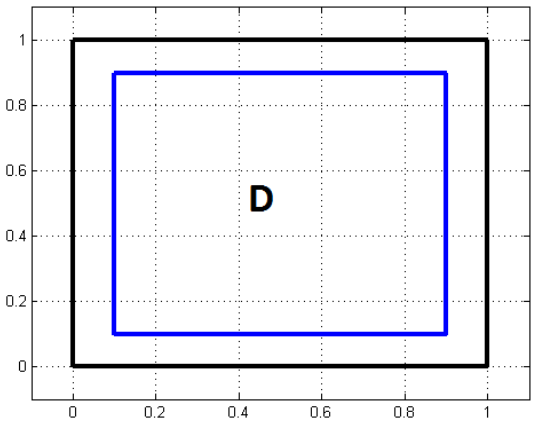


FIGURE 2.9 – Area $D = [0.1; 0.9] \times [0.1; 0.9] \subset [0;1]^2$

Proposition 2.17. *Let the empirical trimmed version of the L-moments issued from the optimal transport be defined by*

$$\hat{\lambda}_\alpha^{(D)} = \int_D Q_n(t) L_\alpha(t) dt \tag{2.78}$$

where $Q_n = T_n \circ Q_0$ is defined in Theorem 2.10. Then

$$\hat{\lambda}_\alpha^{(D)} \xrightarrow{a.s.} \lambda_\alpha^{(D)} \tag{2.79}$$

Proof. We simply perform the same inequality as in Theorem 2.10

$$\begin{aligned}\|\hat{\lambda}_\alpha^{(D)} - \lambda_\alpha^{(D)}\| &\leq \int_D \|T_n(Q_0(t)) - T(Q_0(t))\|dt \\ &\leq \int_\Omega \|T_n(Q_0(t)) - T(Q_0(t))\|dt \\ &\leq \|T_n - T\|_1 \overset{a.s.}{\to} 0\end{aligned}$$

□

2.7.2 Hermite L-moments

Motivation

Until now, L-moments have been defined through an inner product between a transport from the uniform measure on $[0;1]^d$ onto the measure of interest and the elements of an orthogonal basis of polynomials. This definition was motivated by the analogy with the univariate case. In the present multivariate setting, L-moments are not defined as expectations, but namely through Definition 2.6. This allows to consider other source distributions than the uniform one on $[0;1]^d$. It appears that a convenient choice is the standard Gaussian distribution on $\mathbb{R}^d$, which enjoys rotational invariance. In the present context, in contrast with the construction in Section 2.4.3, we do not consider the Gaussian distribution as a reference measure, but directly as the source measure.
As the Legendre polynomials are no longer orthogonal for this distribution, we change the basis as well. Once this basis (P_α) for $\alpha \in \mathbb{N}_*^d$ is chosen, the alternate L-moments are defined by

$$\lambda_\alpha = \int_{\mathbb{R}^d} T(x)P_\alpha(x)d\mathcal{N}_d(x)$$

where T is the monotone transport from $d\mathcal{N}_d$ (the standard multivariate Gaussian measure) onto the target measure.

Definition 2.13. *The univariate orthogonal polynomial basis H_n on the space of functions measurable with respect to $d\mathcal{N}_1$, denoted by $L_2(\mathbb{R}, \mathbb{R}, d\mathcal{N}_1)$ and such that*

$$\langle H_n, H_m\rangle = \int_{\mathbb{R}} H_m(x)H_n(x)d\mathcal{N}_1(x) = \sqrt{2\pi}n!\delta_{nm} \text{ for } n, m \in \mathbb{N} \tag{2.80}$$

are called Hermite polynomials. A multivariate Hermite polynomial is indexed by $\alpha = (i_1, ..., i_d) \in \mathbb{N}^d$:

$$H_\alpha(x) = H_{i_1}(x_1)...H_{i_d}(x_d) \text{ for } x = (x_1, ..., x_d) \in \mathbb{R}^d \tag{2.81}$$

The first univariate Hermite polynomials are

$$\begin{aligned}H_0(x) &= 1 \\ H_1(x) &= x \\ H_2(x) &= x^2 - 1 \\ H_3(x) &= x^3 - 3x \\ H_4(x) &= x^4 - 6x^2 + 3\end{aligned}$$

Proposition 2.18. *The family of multivariate Hermite polynomials $(H_\alpha)_{\alpha\in\mathbb{N}^d}$ is orthogonal and complete in the Hilbert space $L^2(\mathbb{R}^d,\mathbb{R},d\mathcal{N}_d)$ provided with the scalar product :*

$$\langle f,g\rangle = \int_{\mathbb{R}^d} f(x)g(x)d\mathcal{N}_d(x) \tag{2.82}$$

Moreover, for $\alpha=(i_1,...,i_d)$

$$\langle H_\alpha, H_{\alpha'}\rangle = (2\pi)^{d/2} i_1!...i_d! \tag{2.83}$$

Proof. The proof is very similar to the one presented in Lemma 2.5. □

The Hermite L-moments are then defined by

$$\lambda_\alpha = \int_{\mathbb{R}^d} T(x)H_\alpha(x)d\mathcal{N}_d(x). \tag{2.84}$$

Remark 2.33. *Let us note that this definition is not compatible with the L-moments defined by Hosking in the univariate case and for T the monotone transport from $d\mathcal{N}_d$ onto the target measure. Indeed, in that case, Equation 2.84 is written for $\alpha=r\geq 1$*

$$\lambda_r = \int_0^1 Q(u)H_r(Q_{\mathcal{N}_1}(u))du$$

where $Q_{\mathcal{N}}$ and Q respectively are the quantiles of the univariate standard Gaussian and the measure of interest.

Property of invariance/equivariance

The main reason of defining such objects lies in the following property of invariance/equivariance.

Proposition 2.19. *Let X be a random vector in $\mathbb{R}^d$ and $\nabla\varphi$ be the optimal transport from $\mathcal{N}_d$ onto the measure associated to X such that φ is convex. Let denote by $\lambda_\alpha(X)$ the Hermite L-moments of X.*
First, let $\sigma>0, m\in\mathbb{R}^d$. Then

$$\lambda_\alpha(\sigma X+m) = \sigma\lambda_\alpha(X) + m\mathbb{1}_{\alpha=(1...1)}. \tag{2.85}$$

Let us note the L-moment matrix of order two

$$\Lambda_2(X) = \left(\int_{\mathbb{R}^d}(\nabla\varphi)_i(x)H_1(x_j)d\mathcal{N}_d(x)\right)_{i,j=1...d} = \left(\int_{\mathbb{R}^d}(\nabla\varphi)_i(x)x_j d\mathcal{N}_d(x)\right)_{i,j=1...d}. \tag{2.86}$$

Then, if P is an orthogonal matrix (i.e. $PP^T=P^TP=I_d$),

$$\Lambda_2(PX) = P^T\Lambda_2(X)P. \tag{2.87}$$

Remark 2.34. *This second property seems to be particular to the L-moments of degree two.*

Proof. The first part is similar to the proof of Proposition 2.12.
For the second part, let us define the potential $\psi: x\mapsto\varphi(Px)$. Then ψ is convex and if N_d denotes a standard multivariate Gaussian random vector

$$\nabla\psi(N_d) = P^T\nabla\varphi(PN_d) \overset{d}{=} P^TX$$

since N_d is invariant by rotation.
Then, $\nabla\psi$ is the monotone transport associated to $P^T X$. Thus,

$$\begin{aligned}\Lambda_2(PX) &= \int_{\mathbb{R}^d} \nabla\psi(x)x^T d\mathcal{N}_d(x)\\ &= P^T \int_{\mathbb{R}^d} \nabla\varphi(Px)x^T d\mathcal{N}_d(x)\\ &= P^T \int_{\mathbb{R}^d} \nabla\varphi(y)(P^{-1}y)^T d\mathcal{N}_d(y)\\ &= P^T\Lambda_2(X)P\end{aligned}$$

□

Applications for linear combinations of independent variables

The previous property is well adapted for the study of linear combinations of independent variables $Z_1, ..., Z_d$.
We suppose that each variable Z_i is normalized by its second L-moment i.e. $\lambda_2(Z_i) = 1$.
Let us recall from Example 2.22 that if $(e_1, ..., e_d)$ is an orthonormal basis of $\mathbb{R}^d$, $(b_1, ..., b_d)$ the canonical basis, then the following potential

$$\varphi(x) = \sum_{i=1}^{d} \sigma_i\varphi_i(x^T e_i) \tag{2.88}$$

is convex, with each function φ_i convex and $a_i > 0$. If we denote $P := \sum_{i=1}^d e_i b_i^T$ and $D := \sum_{i=1}^d \sigma_i b_i b_i^T$, then

$$\nabla\varphi(x) = P^T D \begin{pmatrix} \varphi_1'(x^T e_1) \\ \vdots \\ \varphi_d'(x^T e_d) \end{pmatrix}. \tag{2.89}$$

If, for each i, we note $\varphi_i' := Q_i \circ \mathcal{N}_d$ with Q_i the quantile of Z_i

$$Y \overset{d}{=} P^T DZ = P^T D \begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix}. \tag{2.90}$$

Since we have $\Lambda_2(Z) = I_d$, we deduce from Proposition 2.19 that

$$\Lambda_2(Y) = PDP^T. \tag{2.91}$$

Let us remark that the covariance of Y is given by

$$Cov(Y) = P^T D \begin{pmatrix} Cov(Z_1) & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & Cov(Z_d) \end{pmatrix} DP. \tag{2.92}$$

The covariance and the Hermite L-moment matrix of degree 2 share the same rotation structure for this family of distributions. A principal component analysis can then be done on either matrix for example. Likewise, we can perform a straightforward estimation of P and D thanks to a method based on the L-moments.

$Tr(\Lambda_2)$ and $\det(\Lambda_2)$ are clearly invariant with respect to the rotation matrix P. Furthermore, we compute some Hermite L-moments of degree 3 in order to identify more invariants of the distribution of Y ; this would lead to specific plug-iin estimators for either P or D. Let $\lambda_r^{(H)}(Z_i)$ be the r-th univariate Hermite L-moment of Z_i i.e.

$$\lambda_r^{(H)}(Z_i) = \int_{\mathbb{R}} Q_i \circ \mathcal{N}_1(x) H_r(x) d\mathcal{N}_1(x) \text{ with } Q_i \text{ the quantile of } Z_i$$

Lemma 2.11. *If we denote by Λ_3 the matrix*

$$\Lambda_3(Y) = \left(\int_{\mathbb{R}^d} (\nabla\varphi)_i(x)(x_j^2 - 1) d\mathcal{N}_d(x) \right)_{i,j=1\ldots d}$$

and if Y is a linear combination of independent variables

$$\Lambda_3(Y) = PD\left(P_{ji}^2 (\lambda_3^{(H)}(Z_i) - \lambda_1^{(H)}(Z_i)) \right)_{i,j=1\ldots d}$$

where $P = (P_{ij})_{i,j=1\ldots d}$.

Proof. Let us recall that the optimal transport associated to Y is

$$\nabla\varphi(x) = PD \begin{pmatrix} T_1(x^T e_1) \\ \vdots \\ T_d(x^T e_d) \end{pmatrix}$$

with $T_i = Q_i \circ \mathcal{N}_1$. Then $\Lambda_3 = PDM$ with the (i,j)-th element of M equal to

$$\begin{aligned} M_{ij} &= \int_{\mathbb{R}^d} T_i(b_i^T P^T x) x_j^2 d\mathcal{N}_d(x) - \lambda_1(Z_i) \\ &= \int_{\mathbb{R}^d} T_i(y_i)(b_j^T P y)^2 d\mathcal{N}_d(y) - \lambda_1(Z_i) \end{aligned}$$

Now let $e'_j = P^T b_j$. If e'_j and b_i are collinear, $b_i^T P^T b_j = 1$ then

$$M_{ij} = \int_{\mathbb{R}} T_i(y)(b_i^T P^T b_j y)^2 d\mathcal{N}_1(y) - \lambda_1(Z_i) = \lambda_3^{(H)}(Z_i) - \lambda_1(Z_i).$$

Otherwise, let note $a_1 = b_i^T e'_j = b_i^T P^T b_j$ and $a_2 = \sqrt{1 - a_1^2}$. If we complete e'_j and b_i with $d-2$ orthonormal vector, we can produce the change of variable corresponding to his new basis i.e.

$$\begin{aligned} M_{ij} &= \int_{\mathbb{R}^2} T_i(y'_1)(a_1 y'_1 + a_2 y'_2)^2 d\mathcal{N}_1(y'_1) d\mathcal{N}_1(y'_2) - \lambda_1(Z_i) \\ &= a_1^2 \int_{\mathbb{R}} T_i(y'_1) y_1^2 d\mathcal{N}_1(y'_1) + a_2^2 \int_{\mathbb{R}} T_i(y'_1) d\mathcal{N}_1(y'_1) - \lambda_1(Z_i) \\ &= (b_i^T P^T b_j)^2 \lambda_3^{(H)}(Z_i) - (b_i^T P^T b_j)^2 \lambda_1^{(H)}(Z_i) \end{aligned}$$

which concludes the proof. □

Consider the case $d = 2$. Then, let $L := diag(\lambda_3^{(H)}(Z_1) - \lambda_1^{(H)}(Z_2))$ and $D = diag(\sigma_1, \sigma_2)$. By Lemma 2.11, it holds

$$\begin{cases} \Lambda_3 &= PDL(P.P) \\ \Lambda_2\Lambda_3 &= PD^2L(P.P) \\ \Lambda_2^{-1}\Lambda_3 &= PL(P.P) \end{cases}$$

where $P.P$ denotes the Hadamard product between P and P. We remark that $\Lambda_2^{-1}\Lambda_3$ is invariant with respect to D. Furthermore, if $\sigma_1 \neq \sigma_2$, we can define a, b such that

$$\begin{pmatrix} a \\ b \end{pmatrix} = -\begin{pmatrix} \sigma_1 & 1 \\ \sigma_2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \end{pmatrix}$$

Then $\Lambda_2\Lambda_3 + a\Lambda_3 + b\Lambda_2^{-1}\Lambda_3 = 0$.

Estimation of Hermite L-moments

Let $x_1, ..., x_n \in \mathbb{R}^d$ be n iid realizations of a random vector X (with measure ν). The estimation of Hermite L-moments uses the estimation of a monotone transport presented in Section 2.6.2. If T_n is such a transport from $d\mathcal{N}_d$ onto $\nu_n = \sum_{i=1}^n \delta_{x_i}$, the estimation of the α-th Hermite L-moment is

$$\hat{\lambda}_\alpha = \int_{\mathbb{R}^d} T_n(x) H_\alpha(x) d\mathcal{N}_d(x) = \sum_{i=1}^n \left(\int_{W_i(T_n)} H_\alpha(x) d\mathcal{N}_d(x) \right) x_i \tag{2.93}$$

with the notations of Section 2.6.2 i.e.

$$W_i(T_n) = \left\{ x \in \mathbb{R}^d \text{ s.t. } T_n(x) = x_i \right\}$$

Theorem 2.11. *Let ν satisfy $\int \|x\| d\nu(x) < +\infty$. Then, we have for $\alpha \in \mathbb{N}_*^d$.*

$$\hat{\lambda}_\alpha = \int_{\mathbb{R}^d} T_n(x) H_\alpha(x) d\mathcal{N}_d(x) \overset{a.s.}{\to} \lambda_\alpha = \int_{\mathbb{R}^d} T(x) H_\alpha(x) d\mathcal{N}_d(x) \tag{2.94}$$

Proof. The proof is very similar to the proof of Theorem 2.10 □

Numerical applications

We will present some numerical results for the estimation of L-moments and Hermite L-moments issued from the monotone transport. For that purpose, we simulate a linear combination of independent vectors in $\mathbb{R}^2$

$$Y = P \begin{pmatrix} \sigma_1 Z_1 \\ \sigma_2 Z_2 \end{pmatrix}$$

with

$$P = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$$

and Z_1, Z_2 are drawn from a symmetrical Weibull distribution ϵW_ν where ϵ is a Rademacher random variable ($\epsilon = 2B - 1$ with B is a Bernoulli a parameter $1/2$) and W_ν is a Weibull of shape parameter ν and scale parameter 1. The density of W_ν is given by

$$f_\theta(x) = 8\nu(8x)^{\nu-1} e^{-(8x)^\nu}.$$

We perform $N = 100$ estimations of the second L-moment matrix Λ_2, the second Hermite L-moment matrix $\Lambda_2^{(H)}$ and the covariance matrix Σ for a sample of size $n = 30$ or 100. We present the results in Table 2.1 through the following features

- The mean of the different estimates

		$n = 30$			$n = 100$		
Parameter	True Value	Mean	Median	CV	Mean	Median	CV
$\Lambda_{2,11}$	0.38	0.28	0.27	0.30	0.38	0.37	0.18
$\Lambda_{2,12}$	0.19	0.14	0.13	0.65	0.20	0.20	0.33
$\Lambda_{2,11}^{(H)}$	0.66	0.5	0.48	0.31	0.70	0.68	0.19
$\Lambda_{2,12}^{(H)}$	0.33	0.25	0.24	0.67	0.38	0.37	0.34
Σ_{11}	0.69	0.70	0.48	1.23	0.69	0.59	0.55
Σ_{12}	0.55	0.55	0.29	1.62	0.54	0.47	0.67

TABLE 2.1 – Second L-moments and covariance numerical results for $\nu = 0.5$

- The median of the different estimates
- The coefficient of variation of the estimates $\hat{\theta}_1, ..., \hat{\theta}_N$ (for an arbitrary parameter θ)

$$CV = \frac{\left(\sum_{i=1}^{N}\left(\theta_i - \frac{1}{N}\sum_{i=1}^{N}\theta_i\right)^2\right)^{1/2}}{\frac{1}{N}\sum_{i=1}^{N}\theta_i}$$

Table 2.1 illustrates the fact that the L-moment estimator are more stable than classical covariance estimates but more biased for heavy-tailed distributions. The effects should be even more visible for moments of higher order. However, our sampled L-moments introduces a bias for small n contrary to classical empirical covariance.

2.8 Appendix

2.8.1 Proof of Theorem 2.6.2

We adapt the proof of Gu et al. [58] to the case of an absolutely continuous probability measure μ defined on $\Omega \subset \mathbb{R}^d$. We do not assume the compactness of Ω.
The proof is divided into four steps

- First, we show that the set $H = \{h \in \mathbb{R}^n \text{ s.t. } vol(W_i(h) \cap \Omega) > 0 \text{ for all i}\}$ is a non-void open convex set
- Secondly, we show that $E_0(h) = \int_\Omega \phi_h(x)d\mu(x)$ is a C^1-smooth convex function on H so that $\frac{\partial E_0(h)}{\partial h_i} = \int_{W_i(h)\cap\Omega} d\mu(x)$
- In the third step, we show that $E_0(h)$ is strictly convex on $H_0^{(n)} = H \cap \{h \in \mathbb{R}^n \text{ s.t. } \sum_{i=1}^n h_i = 0\}$
- Finally, we will prove that $\nabla\phi_h$ is a monotone transport

Convexity of H

Let denote by $H_i = \{h \in \mathbb{R}^n \text{ s.t. } vol(W_i(h) \cap \Omega) > 0\}$. We can remark that the condition $vol(W_i(h) \cap \Omega) > 0$ is the same as assuming that $W_i(h) \cap \Omega$ contains a non-empty open set in $\mathbb{R}^d$.
Furthermore, as $x_1, ..., x_n$ are distinct, if $int(W_i(h)) \neq \emptyset$, then $int(W_i(h)) = \{u \in \mathbb{R}^d \text{ s.t. } u.x_i + h_i > \max_{j\neq i} u.x_j + h_j\}$ (Prop 2.2(a) of Gu et al. [58]). It follows that

$$H_i = \left\{h \in \mathbb{R}^n \text{ s.t. there exists } u \in \Omega \text{ so that } u.x_i + h_i > \max_{j\neq i} u.x_j + h_j\right\}.$$

We prove now that, for any i, H_i is convex which implies that $H = \cap_{i=1}^n H_i$ is convex. Let $\alpha, \beta \in H_i$ and $0 \leq t \leq 1$. Then there exists $v_1, v_2 \in \Omega$ such that $v_1.x_i + \alpha_i > v_1.x_j + \alpha_j$ and $v_2.x_i + \beta_i > v_2.x_j + \beta_j$ for $j \neq i$. Then

$$(tv_1 + (1-t)v_2).x_i + (t\alpha_i + (1-t)\beta_i) > (tv_1 + (1-t)v_2).x_j + (t\alpha_j + (1-t)\beta_j) \text{ for all } j \neq i$$

i.e. $tv_1 + (1-t)v_2 \in H_i$ i.e. H_i is convex. As H_i is non empty since we can take an $h \in \mathbb{R}^n$ such that h_i is as large as needed, we thus have proved that H is an open convex set. Furthermore, as $vol(\Omega) > 0$, there exists a cube included in Ω. We can then translate and rescale the Voronoi cells by the method given in Proposition 2.16 in order to prove that $H \neq \emptyset$.

Convexity of $h \mapsto E_0(h)$ and the expression of its gradient

Let us recall that $E_0(h) = \int_\Omega \phi_h(u) d\mu(u)$ with

$$\phi_h(u) = \max_{1 \leq i \leq n} \{u.x_i + h_i\} \text{ for } u \in \Omega$$

Since functions $(u, h) \mapsto u.x_i + h_i$ are linear, it follows that $(u, h) \mapsto \max_i u.x_i + h_i$ is convex in $\Omega \times \mathbb{R}^n$. Furthermore $d\mu$ is a positive measure. We then have that E_0 is convex in $\mathbb{R}^n$.
Now let $h, d \in \mathbb{R}^n$ and $t > 0$. We consider the Gateaux derivative of E_0

$$\frac{E_0(h+td) - E_0(h)}{t} = \int_\Omega \frac{\phi_{h+td}(u) - \phi_h(u)}{t} d\mu(u).$$

Since ϕ_h is piecewise linear, for almost every $u \in \Omega$, there exists i such that $u \in int(W_i(h))$. Let us choose such a u. Then, clearly, for t small enough $\phi_{h+td}(u) = u.x_i + h_i + td_i$ i.e.

$$\frac{\phi_{h+td}(u) - \phi_h(u)}{t} \to_{t \to 0} d_i.$$

Furthermore, if we take an arbitrary $t > 0$, there exists j such that $\phi_{h+td}(u) = u.x_j + h_j + td_j$. Then

$$\frac{|\phi_{h+td}(u) - \phi_h(u)|}{t} = \frac{|u.(x_i - x_j) + h_i - h_j - td_j|}{t} \leq \frac{\|u\|\|x_i - x_j\| + |h_i - h_j - td_j||}{t}.$$

As $d\mu$ is a probability measure of finite expectation, $u \mapsto \frac{\|u\|\|x_i - x_j\| + |h_i - h_j - td_j||}{t}$ is $d\mu$-measurable. Thus, by the dominated convergence theorem,

$$\frac{E_0(h+td) - E_0(h)}{t} \to_{t \to 0} \sum_{i=1}^n d_i \int_{W_i(h) \cap \Omega} d\mu(u)$$

i.e. E_0 is Gateaux differentiable and $\frac{\partial E_0}{\partial h_i} = \int_{W_i(h) \cap \Omega} d\mu(u)$.
This show furthermore that for some $a \in \mathbb{R}^n$, E_0 could be written

$$E_0(h) = \int_a^h \sum_{i=1}^n \int_{W_i(h) \cap \Omega} d\mu(u) dh_i.$$

Strict convexity on $H_0^{(n)}$

Let $w_i(h) = \frac{\partial E_0}{\partial h_i} = \int_{W_i(h)\cap\Omega} d\mu(u)$. Then $\sum_{i=1}^n w_i(h) = \int_\Omega d\mu(u) = 1$.

Lemma 2.12. *$h \mapsto w_i(h)$ is a differentiable function and if $W_i(h) \cap \Omega$ and $W_j(h) \cap \Omega$ share a face F with codimension 1. Then*

$$\frac{\partial w_i(h)}{\partial h_j} = -\frac{1}{\|x_i - x_j\|}\int_F d\mu_F(u) \text{ for } j \neq i \tag{2.95}$$

Otherwise

$$\frac{\partial w_i(h)}{\partial h_j} = 0 \text{ for } j \neq i. \tag{2.96}$$

Proof. It is the adaptation of the proof of Gu et al. [58] by replacing the compactness assumption by the hypothesis that $d\mu$ is a probability measure. The use of dominated convergence theorem remains unchanged. □

We have proved that E_0 is twice differentiable and we note the Hessian matrix of E_0

$$Hess(E_0) = [a_{ij}]_{i,j=1...n} = \left[\frac{\partial^2 E_0}{\partial h_i \partial h_j}\right]_{i,j=1...n} = \left[\frac{\partial w_i(h)}{\partial h_j}\right]_{i,j=1...n}.$$

Then

$$\sum_{i=1}^n \frac{\partial^2 E_0}{\partial h_i \partial h_j} = \frac{\partial}{\partial h_j} 1 = 0$$

i.e. $(1, ..., 1) \in Ker(Hess(E_0))$. Furthermore, this equality combined with Lemma 2.12 shows that $Hess(E_0)$ is diagonally dominant with positive diagonal entries i.e.

$$a_{ii} = Hess(E_0)_{ii} = -\sum_{j\neq i} a_{ij} \geq 0.$$

As $Hess(E_0)$ is Hermitian, $Hess(E_0)$ is positive semidefinite. It remains to show that (1...1) is the only member of its kernel.
Let $y \in Ker(Hess(E_0))$. Let us assume without loss of generality that $y_1 = \max_{i=1...n} |y_i| > 0$. Then if we combine the two equalities

$$a_{11}y_1 = -\sum_{j=2}^n a_{1j}y_j$$

and

$$a_{11} = -\sum_{j=2}^n a_{1j}$$

we get

$$\sum_{j=2}^n a_{1j}(y_j - y_1) = 0$$

As $a_{1j} \leq 0$ and $y_j \leq y_1$, either $a_{1j} = 0$ either $y_j = y_1$.
Since Ω is a convex domain and each $W_k(h)$ as well, there exists a rearrangement of $(1, ..., n)$, denoted by $i_1, ..., i_n$, such that $W_{i_j} \cap \Omega$ and $W_{i_{j+1}} \cap \Omega$ share a codimension-1 face for each j. We can again assume without loss of generality that $i_1 = 1$. Then by iteration, we find that $y_{i_{j+1}} = y_1$ since $a_{i_j i_{j+1}} < 0$ for any j.
It follows that $y = y_1(1....1)$ i.e. $\dim(Ker(Hess(E_0)) = 0$.

$\nabla\phi_{h^*}$ is an optimal transport map

We produce here the proof of Aurenhammer et al. [9] in order to prove that $\nabla\phi_{h^*}$ minimizes the quadratic transport cost.
Let first remark that the quadratic transport of $\nabla\phi_{h^*}$ is

$$\sum_{i=1}^{n}\int_{W_i(h^*)}\|x_i-u\|^2 d\mu(u).$$

From Lemma 2.8, $\cup_{i=1}^{n}W_i(h^*)$ is the power diagram associated to $(x_1, w_1 = -\|x_1\|^2 - 2h_1^*), ..., (x_n, w_n = -\|x_n\|^2 - 2h_n^*)$. Suppose that $(V_1, ..., V_n)$ is any partition of $\mathbb{R}^d$ such that

$$\int_{V_i} d\mu(u) = \int_{W_i(h^*)} d\mu(u) = \frac{1}{n} \text{ for any } i = 1...n$$

By definition of the power diagram, we get

$$\sum_{i=1}^{n}\int_{W_i(h^*)}(\|x_i-u\|^2+w_i)d\mu(u) \leq \sum_{i=1}^{n}\int_{V_i}(\|x_i-u\|^2+w_i)d\mu(u)$$

i.e.

$$\sum_{i=1}^{n}\int_{W_i(h^*)}\|x_i-u\|^2 d\mu(u) \leq \sum_{i=1}^{n}\int_{V_i}\|x_i-u\|^2 d\mu(u).$$

This shows that $\nabla\phi_{h^*}$ minimized the quadratic cost.
Alternatively, we could mention that as $\nabla\phi_{h^*}$ is a gradient of a convex function which transports μ onto ν_n, we get the result above by Proposition 2.10. However, the proof given above makes this result explicit.

Chapitre 3

M-estimators of the scatter matrix for stationary and non-stationary elliptical distributions

3.1 Introduction

3.1.1 Motivation and notations

Non-Gaussian models of strong clutters such as ground or sea clutters are used in the field of radar processing. The family of complex elliptically symmetric distributions [83] (which contains a lot of classical distributions such as multivariate Gaussian, multivariate Cauchy distributions and multivariate K-distributions) is a useful generalization of Gaussian random vectors, inheriting of similar shape and location parameters.
We consider the case when the location parameter is zero.Denote $X = (X_1, ..., X_d)^T \in \mathbb{C}^d$ a random vector with complex elliptical symmetric (CE) distribution (to be defined hereafter). A CE models the "angle" of X with $\frac{X}{\|X\|}$ distributed on the unit sphere on $\mathbb{C}^d$ and the amplitude of X, namely $\|\Sigma^{-1/2}X\|$ where Σ is the scatter matrix. Assuming that $x_1, .., x_N$ is an iid sample with CE distribution on $\mathbb{C}^d$, the main focus of our study lies in the estimation of the scatter matrix Σ of the underlying distribution. Observe $(x_{11}, ..., x_{1d})^T, ..., (x_{N1}, ..., x_{Nd})^T$ N iid realizations of the vector $(X_1, ..., X_d)^T$. Within this framework, we consider two kinds of robustness concepts for the estimation of the scatter matrix :

- (R1) a robustness with respect to the distribution of the amplitude which is often heavy-tailed
- (R2) a robustness with respect to contamination in the observed sample

First of all, we consider stationary samples where stationarity is defined as the second order one. This assumption adds a Toeplitz structure constraint for the scatter matrix Σ. Taking into account the Toeplitz intrinsic structure of the scatter matrix can be performed by solving the equation 3.2 hereafter in the space of Toeplitz matrices (see [86]). Our approach is slightly different with respect to the above one.
The Toeplitz structure allows us to split the estimation of the matrix Σ of size $d \times d$ into d estimations of Toeplitz matrices of size 2×2. This splitting corresponds to the so-called "Burg technique" [31]. Indeed, instead of estimating the covariance of the raw sample $x_1, ..., x_N \in \mathbb{C}^d$, we iteratively define second-order samples in $\mathbb{C}^2$ whose theoretical covariance can be expressed in function of Σ.
This technique was originally proposed in the context of stationary Gaussian autoregres-

sive time series. The sample $x_1, ..., x_N$ can be viewed as the collection of N traces of such a time series. The parallel between a time series and its trace is often implicit in the signal processing literature. For this reason, we will refer at this trace as autoregressive vector. Moreover, if we consider X as the trace of an autoregressive process of order $M < d-1$, we add more structure on the matrix Σ than the Toeplitz one. Actually, given the autocovariance $\mathbb{E}[X_1\overline{X_k}]$ for $k = 1...M$ with $M \leq d-1$, it is well known that the maximum entropy model pertaining to the vector $X = (X_1, ..., X_d)^T$ in $\mathbb{C}^d$ results as the complex Gaussian distribution in $\mathbb{C}^d$, whose covariance coincides with the autoregressive autocovariance of size $d \times d$ (see [31][89]).
We propose here to adapt these techniques for non-Gaussian scale mixtures of autoregressive vectors, a big subfamily of the class of elliptical distribution with the autoregressive intrinsic structure.

In the general case of elliptical distributions with a known center (that we will suppose to be zero), Maronna proposed Huber-type robust M-estimators of the scatter matrix Σ satisfying the equation :

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N} u(x_i^+\Sigma^{-1}x_i)x_i x_i^+, \tag{3.1}$$

see [78].
The function u has to satisfy some conditions for the estimator to be defined and consistent. A major drawback of these estimators is their non-invariance with respect to the distribution of the amplitude. For this sake, Tyler [104] (and extended by Pascal et al. [88] in the complex case) proposed another estimator satisfying

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N} \frac{x_i x_i^+}{x_i^+\Sigma^{-1}x_i}. \tag{3.2}$$

The function $u(x) = \frac{1}{x}$ does not satisfy the conditions of Maronna but Tyler has shown that this last estimator is well defined and consistent. It was furthermore shown to be a maximum likelihood estimator for normalized samples $\frac{x_1}{\|x_1\|}, .., \frac{x_N}{\|x_N\|}$ (often called multivariate signs). It is obvious that Tyler's estimator satisfies the first robustness (R1). Unfortunately, the robustness with respect to inhomogeneous distribution (R2) may not hold.
Hallin et al. [59] refine Tyler's estimator by adding the information provided by the rank r_i of $d_i = x_i^+\Sigma^{-1}x_i$ among $d_1, ..., d_N$. The ranks $r_1, ..., r_N$ are also invariant with respect to the amplitude. Hallin et al. showed that their estimator is optimal in a semi-parametric sense that will not be developed here. On empirical basis, this estimator performs greater improvements for elliptical distributions with light tails than heavy tails.
We propose here to adapt Huber M-estimators used by Maronna (see for example Huber and Ronchetti [66]) for the distribution of the normalized samples $\frac{x_1}{\|x_1\|}, .., \frac{x_N}{\|x_N\|}$. The random vector $\frac{X}{\|X\|}$is known to have a so-called Angular Central Gaussian (ACG) [105]. We combine the robustness of the Huber approach with the invariance with respect to the distribution of the amplitude. We will give herein some new consistence results and the expression of the influence functions of the scatter matrix.

We may consider a Burg type technique for the estimation of the scatter matrix ; this is achieved through an iterative procedure. Unfortunately, this method cannot be combined with a robust procedure at each step ; see section 3.7.1. We then propose a geometrical method consisting in computing the median of autoregressive models estimated for

		Model		
		Elliptical	Stationary elliptical	Autoregressive elliptical
Asked Feature	Independence with respect to the amplitude (R1)	Tyler's Fixed Point (3.2)	Stationary Fixed Point [86] Normalized Burg(3.23) Log Burg (3.26) Elliptical Burg (3.29)	Normalized Burg(3.23) Log Burg (3.26) Elliptical Burg (3.29)
	Robustness to contamination (R2)	Maronna estimator [78]		
	Robustness and independence (R1)+(R2)	ACG non-normalized (3.53) ACG normalized (3.58)	Geodesic Burg estimators (3.7.2)	

Table 3.1 – Different features proposed by the estimators of the scatter matrix Σ (the new estimators proposed in this chapter are underlined)

subsamples of $x_1, ..., x_N$ [6][10][12]. The known robustness of the median with respect to outliers will be illustrated.

Most of the estimators proposed are M-estimators that is to say that they are the solution to a minimization (or maximization) problem. Euclidean steepest descent fails to converge quickly. We shall propose some Riemannian descent algorithms for a natural metric defined on the parameter space.

The conjugate transpose of a vector or a matrix x would be denoted by x^+ and the real part of a complex z is $\Re(z)$.

3.1.2 Models

Elliptical distributions

We consider N samples $x_1, ..., x_N$ coming from a complex elliptical distribution $CE(\mu, \Sigma, \phi)$; hence each of the x_i's is a vector in $\mathbb{C}^d$. The distribution $CE(\mu, \Sigma, \phi)$ is semi parametric and can be defined as follows.

Definition 3.14. *A complex elliptical random vector $X \in \mathbb{C}^d$ is defined by its characteristic function, which satisfies :*

$$\Phi_X(t) = \mathbb{E}\left[e^{i\Re(t^+X)}\right] = \exp\left(i\Re(t^+\mu)\right)\phi(t^+\Sigma t) \tag{3.3}$$

for $t \in \mathbb{C}^d$, where

- *$\phi : \mathbb{R}^+ \to \mathbb{R}$ is called the characteristic generator*
- *Σ is a $d \times d$ complex positive semi-definite Hermitian matrix representing the covariance of the distribution up to a multiplicative constant*
- *$\mu \in \mathbb{C}^d$ is the expectation (whenever it exists) .*

For the identifiability of the model, we will specify :

$$Tr(\Sigma) = d \tag{3.4}$$

This family is closely related to the real symmetric elliptical parametric family proposed by Cambanis et al. [32].

An important reformulation of the above definition is given by the following characterization

Proposition 3.20. *A random vector $X \in \mathbb{C}^d$ follows a complex elliptical distribution $CE(\mu, \Sigma, \phi)$ with $rank(\Sigma) = k$ if and only if it admits the stochastic representation*

$$X \stackrel{d}{=} \mu + \sqrt{Q} A U \tag{3.5}$$

where the random scalar $R = \sqrt{Q} > 0$ is independent of the uniform random vector U on the sphere $S_k = \left\{ x \in \mathbb{C}^k \text{ s.t. } \|x\|_2 = 1 \right\}$ and A is a $d \times k$ complex positive semi-definite Hermitian matrix with $rank(A) = k$ and $AA^+ = \Sigma$.

Proof. The proof of this characterization can be built similarly as done in the real case (see Cambanis et al. [32]).
We begin with the result of Schoenberg [95]

Lemma 3.13. *If for $k \geq 1$,*

$$\Phi_k = \left\{ \phi : [0; +\infty) \to \mathbb{R} \text{ such that } t \in \mathbb{R}^k \mapsto \phi(\|t\|^2) \text{ is a characteristic function} \right\}$$

then $\phi \in \Phi_k$ if and only if

$$\phi : u \in \mathbb{R}^k \mapsto \phi(u) = \int_0^{+\infty} \Omega_k(r^2 u) dF(r)$$

for some distribution function F on $[0; +\infty)$ and $t \in \mathbb{R}^k \mapsto \Omega_k(\|t\|^2)$ is the characteristic function of a k-dimensional random vector $U_{\mathbb{R}}^{(k)}$ which is uniformly distributed on the unit sphere in $\mathbb{R}^k$.

We remark also that if $U = U_x + iU_y$ is a random variable uniformly distributed on the unit sphere in $\mathbb{C}^k$, (U_x, U_y) is uniformly distributed on the unit sphere in $\mathbb{R}^{2k}$.

Let us first suppose that $X \stackrel{d}{=} \mu + RAU$. Then, if $t \in \mathbb{C}^k$

$$\begin{aligned}
\phi_X(t) &= \mathbb{E}\left[e^{i\Re(t^+ X)} \right] \\
&= e^{i\Re(t^+ \mu)} \mathbb{E}\left[e^{iR\Re(t^+ AU)} \right] \\
&= e^{i\Re(t^+ \mu)} \int_0^{\infty} \mathbb{E}_U \left[e^{ir\Re((A^+ t)^+ U)} \right] dF(r)
\end{aligned}$$

Furthermore

$$\begin{aligned}
\Psi(v) &= \mathbb{E}\left[e^{i\Re(v^+ U)} \right] \\
&= \mathbb{E}\left[e^{i(v_x U_x - v_y U_y)} \right] \\
&= \Omega_{2k}(\|v_x\|^2 + \|v_y\|^2) \\
&= \Omega_{2k}(\|v\|^2)
\end{aligned}$$

i.e.

$$\begin{aligned}\phi_X(t) &= e^{i\Re(t^+\mu)}\int_0^\infty \Omega_{2k}(r^2 t^+ AA^+ t)dF(r)\\ &= e^{i\Re(t^+\mu)}\phi(t^+\Sigma t)\end{aligned}$$

Secondly, we will suppose that the characteristic function of X is $\phi_X(t) = e^{i\Re(t^+\mu)}\phi(t^+\Sigma t)$. As $rank(A) = k$, we can denote by A^- its left inverse. We have $A^-A = I_k$.
If $Z = A^-(X-\mu)$, its characteristic function is given for $t \in \mathbb{C}^k$

$$\begin{aligned}\phi_Z(t) &= \mathbb{E}\left[e^{i\Re(t^+Z)}\right]\\ &= \phi_X((A^-)^+t)\\ &= \phi(t^+A^-AA^+(A^-)^+t) = \phi(\|t\|^2) = \phi(\|t_x\|^2 + \|t_y\|^2)\end{aligned}$$

As $(t_x, t_y) \mapsto \phi_Z(t_x + it_y) = \mathbb{E}[e^{i(t_x Z_x - t_y Z_y)}]$ is a real characteristic function, we can apply Schoenberg Lemma :

$$\phi_Z(t) = \int_0^\infty \Omega_{2k}(r^2\|t\|^2)dF(r)$$

i.e. there exists a positive random variable R such that $(Z_x, Z_y) \overset{d}{=} RU_{\mathbb{R}}^{(2k)}$ where $U_{\mathbb{R}}^{(2k)}$ is uniformly distributed on the real unit sphere.
We conclude by noting that $Z = Z_x + iZ_y \overset{d}{=} RU$ with U uniformly distributed on the complex unit sphere of $\mathbb{C}^k$. □

Example 3.25. *(Multivariate Gaussian distribution)*
Let R_σ be a random variable with generalized Gamma distribution with density

$$f_\sigma(x) = \frac{x^{2d-1}}{\Gamma(d)\sigma^{2d}}e^{-\frac{x^2}{2\sigma^2}}$$

Let U_d be uniformly distributed over the unit sphere of $\mathbb{C}^d$ and let X be defined by

$$X \overset{d}{=} R_\sigma \Sigma^{1/2} U_d$$

where Σ is a $d \times d$ invertible matrix such that $Tr(\Sigma) = d$. Then, X is a centered multivariate Gaussian of covariance matrix $\sigma\Sigma$ and X is an elliptically-distributed random vector thanks to the characterization given by Proposition 3.20.

Example 3.26. *(Multivariate Weibull distribution)*
Let $W_{\theta,\sigma}$ be a random variable with Weibull distribution with density

$$f_{\theta,\sigma}(x) = \frac{\theta}{\sigma}\left(\frac{x}{\sigma}\right)^{\theta-1} e^{-(x/\sigma)^\theta}$$

Let U_d be uniformly distributed over the unit sphere of $\mathbb{C}^d$ and let X be defined by

$$X \overset{d}{=} W_{\theta,\sigma}\Sigma^{1/2}U_d$$

Then X is an elliptically-distributed random vector.

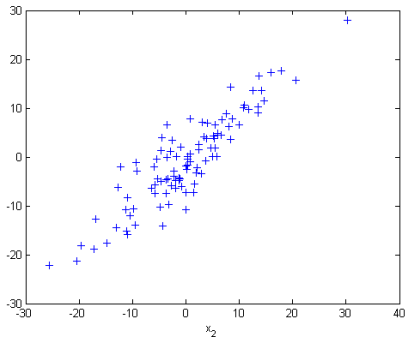


FIGURE 3.1 – A sample of size 100 of a real multivariate Gaussian distribution with $\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$ and $\sigma = 10$

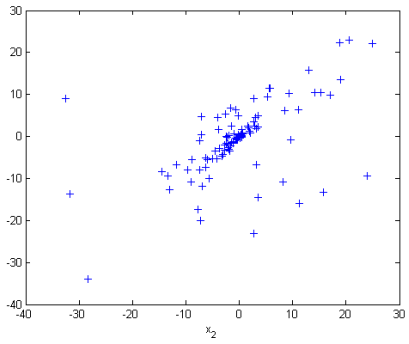


FIGURE 3.2 – A sample of size 100 of a real multivariate Weibull distribution with $\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$, $\theta = 1$ and $\sigma = 10$

Example 3.27. *(Multivariate K-distribution)*
Let $K_{\theta,\sigma}$ be a random variable with K-distribution with density

$$f_{\theta,\sigma}(x) = \frac{2}{x}\left(\frac{\theta x}{\sigma}\right)^{\frac{\theta+1}{2}} \frac{1}{\Gamma(\theta)} K_{\theta-1}\left(2\sqrt{\frac{\theta x}{\sigma}}\right)$$

where K is a modified Bessel function of the second kind

$$K_{\theta-1} = \int_0^{+\infty} \exp(-x\cosh t)\cosh((\theta-1)t)dt$$

Let U_d be uniformly distributed over the unit sphere of $\mathbb{C}^d$ and let X be defined by

$$X \overset{d}{=} K_{\theta,\sigma}\Sigma^{1/2}U_d$$

A K-distributed random variable can be represented by a compound Gaussian variable $K_{\theta,\sigma} \overset{d}{=} G_{\theta,\sigma}R$ where R follows the generalized Gamma distribution of scale parameter 1 and $G_{\theta,\sigma}$ is Gamma-distributed with a shape parameter θ and scale parameter σ.

This distribution is often used for the modelization of sea clutter [55] [67].

For more details on complex elliptical distributions, see Ollila et al. [83].

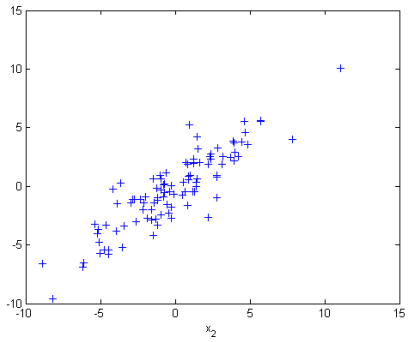


Figure 3.3 – A sample of size 100 of a real multivariate K-distribution with $\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$, $\theta = 1$ and $\sigma = 10$

Complex Angular Gaussian (ACG) distributions

In the following we will only consider the case $\mu = 0$. Moreover, in order to ensure the invariance of the inference with respect to the amplitude R, we will only consider in the following the distribution of the normalized samples $\frac{x_1}{\|x_1\|}, .., \frac{x_N}{\|x_N\|}$. If $\text{rank}(\Sigma) = d$, their common parametric distribution is called angular central Gaussian $ACG(\Sigma)$ whose density is given by

$$f_\Sigma(x) = \frac{\Gamma(d)}{2\pi^d|\Sigma|}(x^+\Sigma^{-1}x)^{-d}\mathbb{1}_{\|x\|=1} \tag{3.6}$$

The parameter Σ is defined up to a multiplicative constant since $f_{a\Sigma}$ does not depend on $a > 0$. For the identifiability of the model, we specify the same constraint than for elliptical models i.e.

$$Tr(\Sigma) = d. \tag{3.7}$$

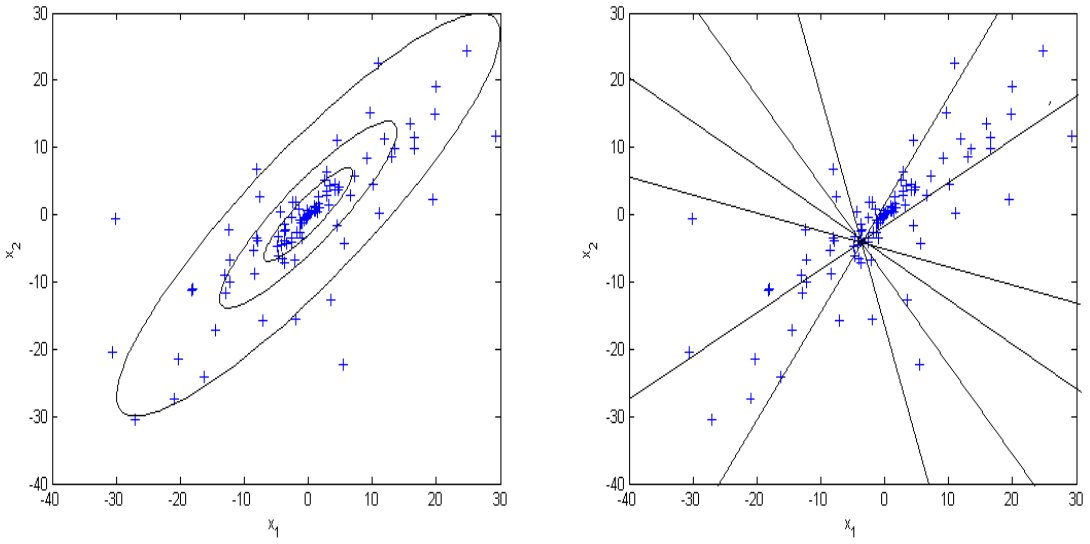
Any estimation of this parameter should assume this constraint.
For a sample $z_1 = \frac{x_1}{\|x_1\|}, ..., z_N = \frac{x_N}{\|x_N\|}$ following this distribution, the maximum likelihood estimate of the scatter matrix Σ corresponds to the Tyler estimator defined by

$$\hat{\Sigma}_T = \sum_{i=1}^{N} \frac{z_i z_i^+}{z_i^+ \hat{\Sigma}_T^{-1} z_i} = \sum_{i=1}^{N} \frac{x_i x_i^+}{x_i^+ \hat{\Sigma}_T^{-1} x_i}. \tag{3.8}$$

The maximum likelihood is known to be efficient but not robust to outliers or contamination which is a common situation in radar.

Remark 3.35. *The statistic $\frac{x}{\|x\|}$ is not sufficient for the scatter matrix in the elliptical model 3.3. We lack some information by considering the observed normalized vectors $(z_1, ..., z_N)$.*

The definition of an outlier is clearly depending upon the distribution under consideration. When considering ACG, an outlier is defined through an angle far away from the principal directions of the scatter matrix Σ as illustrated in Figure 3.4b. Indeed, the norm of the vector x does not matter for the ACG distribution. For elliptical models 3.3, an outlier is some point x far away from the ellipsoids in the sense that $x\Sigma^{-1}x$ takes exceedingly large values ; see figure 3.4a.



(a) Elliptical isodensities (b) ACG isodensities

FIGURE 3.4 – Comparison between the isodensity curves (of respective mass 0.5, 0.75 and 0.95) for the elliptical distribution and the angular complex Gaussian (ACG) distribution for the same 100 realizations of a real elliptical distribution

3.1.3 Considered contamination

We have to face the situation of a heavy contamination while estimating the scatter matrix. In practice, this means that $X_1, ..., X_{N_0}$ are drawn from an elliptical distribution $CE(0, \Sigma_0, \phi)$ and $X_{N_0+1}, ..., X_N$ are outliers. We clearly assume that $N_0 > \frac{N}{2}$. In the context of clutter transition for example, the outliers are drawn from an elliptical distribution with a different scatter matrix Σ_1. However, the context can be more complicated : presence of targets in the sample, etc.
Let $P_{\Sigma_0} := CE(0, \Sigma_0, \phi)$ and Q be the law of the outliers. The samples are then drawn from the gross error model

$$P = (1 - \epsilon)P_{\Sigma_0} + \epsilon Q$$

with $0 \leq \epsilon < 1$. The robust estimation aims to estimate Σ_0 regardless of the form of Q. In our applications, we will turn our attention to outliers laws of the form $Q = P_{\Sigma_1}$ with $\Sigma_1 \neq \Sigma_0$. For example, Σ_1 can represent the Doppler covariance matrix of a rainy cloud whereas Σ_0 represents the ambient ground clutter.

3.2 Stationary elliptical models : scale mixture of autoregressive vectors

Let $X \in \mathbb{C}^d$ be a random variable sampled from a scale mixture of stationary Gaussian autoregressive random vectors. Then X is characterized by the existence of a scalar random variable $\tau \in \mathbb{R}_+$ and a scatter matrix Σ such that :

$$X \overset{d}{=} \tau Y \tag{3.9}$$

where $Y \sim \mathcal{N}_d(0, \Sigma)$ is the trace of a stationary Gaussian autoregressive process (i.e. a Gaussian vector, called speckle, of Toeplitz covariance Σ) independent of τ (called texture).

The structure of Σ is related to the underlying autoregressive process. Note thatΣ is then defined up to a multiplicative constant due to the presence of τ (we can multiply Σ and divide τ by the same positive constant without changing the vector X). We will normalize Σ by assuming $tr(\Sigma) = d$.
As Y is the trace of a stationary Gaussian autoregressive process of order $M \leq d-1$, if we note

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_d \end{pmatrix}$$

there exist $a_1^{(M)}, ..., a_M^{(M)} \in \mathbb{C}$ such that for $1 \leq n \leq d$:

$$Y_n + \sum_{i=1}^{M} a_i^{(M)} Y_{n-i} = b_n \tag{3.10}$$

where b_n is a complex standard Gaussian random variable independent of $Y_{n-1}, ..., Y_{n-M}$ with the convention $Y_{-i} = 0$ for all $i \geq 0$. Equation 3.10 may be seen as a definition for b_n.
We note that X is also the trace of an autoregressive process with dependent non-Gaussian innovations :

$$X_n + \sum_{i=1}^{M} a_i^{(M)} X_{n-i} = \tau b_n \tag{3.11}$$

3.2.1 Burg method applied to Gaussian autoregressive vectors

We first present the Burg method for Gaussian autoregressive vectors. All the definitions which we introduce for the process where Y is the trace remain valid for the process associated to X.
Let us define the autocovariance function γ of the underlying Gaussian autoregressive process. For $t \geq 0$, we have $\gamma(t) = \mathbb{E}[Y_{n+t}\overline{Y_n}]$ for any n. The covariance Σ of Y is then equal to the autocovariance γ of size $d \times d$

$$\Sigma = \begin{pmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(d-1) \\ \overline{\gamma(1)} & \gamma(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \gamma(1) \\ \overline{\gamma(d-1)} & \dots & \overline{\gamma(1)} & \gamma(0) \end{pmatrix}$$

The autocovariance γ is independent of n because of the stationarity of Y. Moreover, the stationarity condition is captured by the Yule-Walker equation :

$$\begin{pmatrix} \gamma(0) & \dots & \gamma(M-1) \\ \overline{\gamma(1)} & \dots & \gamma(M-2) \\ \vdots & \vdots & \vdots \\ \overline{\gamma(M-1)} & \dots & \gamma(0) \end{pmatrix} \begin{pmatrix} a_M^{(M)} \\ \vdots \\ a_1^{(M)} \end{pmatrix} = - \begin{pmatrix} \gamma(M) \\ \vdots \\ \gamma(1) \end{pmatrix}$$

The Levinson algorithm inverts this equation by introducing the successive autoregressive parameters $(a_k^{(m)})_{1 \leq k \leq m}$ of order $1 \leq m \leq M$:

- Initialization : let us define $P_0 = \gamma(0)$ and

$$\begin{cases} \mu_1 := a_1^{(1)} = -\frac{\gamma(1)}{P_0} \\ P_1 := P_0(1 - |\mu_1|^2) \end{cases} \tag{3.12}$$

- for $1 \leq m \leq M-1$

$$\begin{cases} \mu_{m+1} := a_{m+1}^{(m+1)} = -\frac{\gamma(m+1)+\sum_{k=1}^{m} a_k^{(m)}\gamma(m+1-k)}{P_m} \\ P_{m+1} := P_m(1-|\mu_m|^2) \\ \begin{pmatrix} a_1^{(m+1)} \\ \vdots \\ a_m^{(m+1)} \end{pmatrix} = \begin{pmatrix} a_1^{(m)} \\ \vdots \\ a_m^{(m)} \end{pmatrix} + \mu_{m+1} \begin{pmatrix} \overline{a}_m^{(m)} \\ \vdots \\ \overline{a}_1^{(m)} \end{pmatrix} \end{cases} \tag{3.13}$$

This algorithm improves the Gauss-Jordan matrix inversion algorithm which runs in $O(d^3)$ by taking into account the Toeplitz structure. The Levinson algorithm, computed in $O(d^2)$ operations, enhances the role of the parameters $(\mu_m)_{1\leq m\leq M}$, called **reflection (or Verblunsky) parameters**, that are sufficient, together with P_0, in order to describe the autoregressive vector in $\mathbb{C}^d$. The introduction of the parametrization of a Toeplitz matrix Σ through its reflection parameters came back from 1933 by Verblunsky [107]. It was slightly improved few years after Levinson by Trench [102], relaxing the Hermitian constraint implicitly considered by Levinson and Verblunsky.
Instead of estimating the covariance matrix directly from the samples which does not guarantee the Toeplitz constraint, we estimate these reflection parameters satisfying the Toeplitz structure (we will then use the bijection between $(P_0, \mu_1, ..., \mu_M)$ and Σ given by equations (3.12) and (3.13) to recover an estimated covariance).
For this purpose, Burg proposes in the Gaussian framework to minimize an error at each step $1 \leq m \leq M$:

$$U^{(m)} = \sum_{n=m+1}^{d} |f_m(n)|^2 + |b_m(n)|^2 \tag{3.14}$$

with f_m and b_m respectively the "forward" and "backward" errors defined for $m+1 \leq n \leq d$:

$$\begin{cases} f_m(n) := Y_n + \sum_{k=1}^{m} a_k^{(m)} Y_{n-k} \\ b_m(n) := Y_{n-m} + \sum_{k=1}^{m} \overline{a}_k^{(m)} Y_{n-m+k} \end{cases} . \tag{3.15}$$

Note that the definition of the errors is still valid for $m = 0$.

$$\begin{cases} f_0(n) = Y_n \\ b_0(n) = Y_n \end{cases} \text{ for } 1 \leq n \leq d \, . \tag{3.16}$$

Moreover, thanks to the equation (3.13), we can state for $m+2 \leq n \leq d$:

$$\begin{cases} f_{m+1}(n) = f_m(n) + \mu_{m+1} b_m(n-1) \\ b_{m+1}(n) = b_m(n-1) + \overline{\mu_{m+1}} f_m(n) \end{cases} . \tag{3.17}$$

Note that the errors are random variables and that $f_m(n)$ and $b_m(n)$, both depending on $(a_1^{(m)}, ..., a_m^{(m)})$, are not directly observable from Y for $m > 0$. Although Burg introduces the criterion 3.14 as an iterative least square method for autoregressive process of order m (for m going from 1 to M), it can be justified as an approximation of the likelihood of the errors $e_m(n) := \begin{pmatrix} f_m(n) \\ b_m(n-1) \end{pmatrix}$ for $n \in \{m+1, ..., d\}$. Let us first give the moments of e_m by the following Lemma

Lemma 3.14. *If $m \geq 0$ and $m+1 \leq n \leq d$*

$$\begin{cases} \mathbb{E}[|f_m(n)|^2] = \mathbb{E}[|b_m(n-1)|^2] = P_m \\ \mathbb{E}[f_m(n)\overline{b_m(n-1)}|] = -P_m\mu_{m+1} \end{cases} \tag{3.18}$$

Proof. It is an application of the Proposition 1 of [28] applied for $\{k_1, .., k_m\} = \{1, ..., m+1\}$. □

Hence, the vectors $e_m(n)$ are dependent Gaussian vectors of covariance matrix

$$\Sigma_e = P_m \begin{pmatrix} 1 & -\mu_{m+1} \\ -\overline{\mu_{m+1}} & 1 \end{pmatrix}$$

Let us fix m. The Burg criterion 3.14 would correspond to the maximum likelihood of the vectors $(e_m(n))_{n=m+1...d}$ over all possible P_m if the sequence $(e_m(m+1), ..., e_m(d))$ were composed by independent random vectors. Burg's technique lies in the iterative estimation of the correlation $-\mu_{m+1}$ of the coordinates $f_m(n)$ and $b_m(n-1)$.
The estimation of the reflection parameters $(\mu_m)_{1\leq m\leq M}$ consists then in the solution of the minimization of the empirical error, which is, for a sample $x_1, .., x_N$:

$$\hat{U}^{(m+1)} = \frac{1}{N}\sum_{i=1}^{N}\sum_{n=m+2}^{d} |f_{i,m+1}(n)|^2 + |b_{i,m+1}(n)|^2$$

where, for each i, $f_{i,m}$ and $b_{i,m}$ are the observed forward and backward errors for the sample Y_i. Knowing $\mu_1, ..., \mu_m$, we estimate μ_{m+1} by :

$$\begin{aligned}\hat{\mu}_{m+1}^{(gauss)} &= \arg\min_{\mu_{m+1}} \hat{U}^{(m+1)} \\ &= \arg\min_{\mu_{m+1}} \frac{1}{N}\sum_{i=1}^{N}\sum_{n=m+2}^{d}\Big[\,|f_{im}(n) + \mu_{m+1}b_{im}(n-1)|^2 \\ &+ |b_{im}(n-1) + \overline{\mu_{m+1}}f_{im}(n)|^2\Big]\end{aligned}$$

The solution is given by :

$$\hat{\mu}_{m+1}^{(gauss)} = -2\frac{\sum_{i=1}^{N}\sum_{n=m+2}^{d} f_{im}(n)\overline{b_{im}(n-1)}}{\sum_{i=1}^{N}\sum_{n=m+2}^{d} |f_{im}(n)|^2 + |b_{im}(n-1)|^2} \tag{3.19}$$

3.2.2 Burg method for non-Gaussian vectors

We now consider the autoregressive vector X. The forward and backward errors are defined in analogy with equation (3.15). We see that the estimator defined by equation (3.19) applied on the sample $(x_1, ..., x_N)$ will suffer from the disparity of the realizations of the scalar part $(\tau_1, ..., \tau_N)$. This lead us to adapt the method by considering a different criterion from 3.14 for the error.
It is customary to define the error criterion to be minimized by the term "energy". We propose three different choices for the energy.

Normalized Energy

The first idea could be to consider an error which is normalized with respect to τ :

$$U^{(m+1)} = \sum_{n=m+2}^{d} \frac{|f_{m+1}(n)|^2 + |b_{m+1}(n)|^2}{|f_m(n)|^2 + |b_m(n-1)|^2}. \tag{3.20}$$

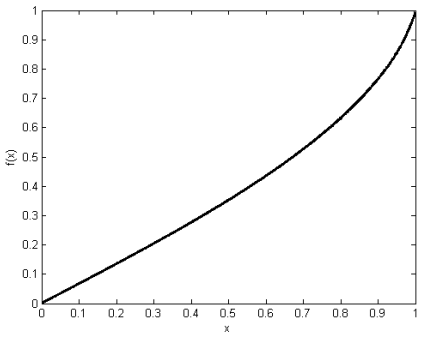


FIGURE 3.5 – Bias function B_1

The minimum of the empirical version of the previous error is then :

$$\hat{\mu}_{m+1} = -\frac{2}{N(d-m-1)} \sum_{i=1}^{N} \sum_{n=m+2}^{d} \frac{\overline{b_{i,m}(n-1)} f_{i,m}(n)}{|f_{i,m}(n)|^2 + |b_{i,m}(n-1)|^2}. \tag{3.21}$$

The drawback is that $\hat{\mu}_{m+1}$ is not consistent. Lemma 3.18 gives us the expression of the moments of $e_m(n)$. It follows that the asymptotic multiplicative bias is tractable :

Proposition 3.21. *For* $1 \leq m \leq M$

$$\hat{\mu}_m \overset{a.e.}{\to} B_1(|\mu_m|)\frac{\mu_m}{|\mu_m|}$$

with B_1 *defined for* $x > 0$ *by :*

$$B_1(x) = \frac{1-x^2}{x}\left(\frac{\log(1-x) - \log(1+x)}{2x} + \frac{1}{1-x^2}\right). \tag{3.22}$$

Proof. This is an application of Theorem 1 of [16] applied to the vector $\begin{pmatrix} f_m(n) \\ b_m(n-1) \end{pmatrix}$. We apply the law of large numbers for the empirical sum $\hat{\mu}_m$. □

The consistent version of (3.21) is then obtained through :

$$\hat{\mu}_m^{(u)} = B_1^{-1}(|\hat{\mu}_m|)\frac{\hat{\mu}_m}{|\hat{\mu}_m|}, \tag{3.23}$$

where B_1^{-1} is not explicit but can be pre-computed for a gain of time.

Remark 3.36. *Another intuitive normalized error could have been :*

$$U^{(m+1)} = \frac{\sum_{n=m+2}^{d} |f_{m+1}(n)|^2 + |b_{m+1}(n)|^2}{\sum_{n=m+2}^{d} |f_m(n)|^2 + |b_m(n-1)|^2} \tag{3.24}$$

which leads to the following estimator

$$\hat{\mu}_{alt,m+1} = -\frac{2}{N(d-m-1)} \sum_{i=1}^{N} \frac{\sum_{n=m+2}^{d} \overline{b_{i,m}(n-1)} f_{i,m}(n)}{\sum_{n=m+2}^{d} |f_{i,m}(n)|^2 + |b_{i,m}(n-1)|^2}.$$

The drawback of this estimator is the non-consistency if m *is fixed and* $N \to \infty$*. Despite its theoretical disadvantage, it has good behavior in practice especially for high values of* m *because it is an average of the Gaussian estimators of* μ_{m+1} *for each* $1 \leq i \leq N$*.*

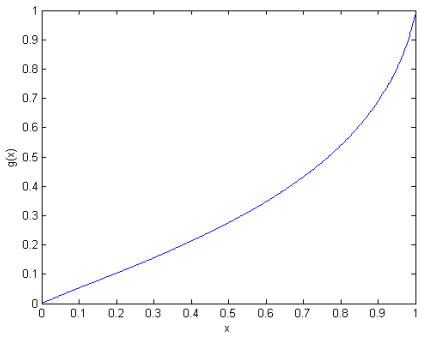


Figure 3.6 – Bias function g of the Log Burg estimator

Logarithmic Energy

We now introduce a logarithmic type of energy. It shares with the previous one the property of independence with respect to the amplitude τ. Let us define :

$$U^{(m)} = \sum_{n=m+1}^{d} -\log\left(|f_m(n)|^2 + |b_m(n)|^2\right). \tag{3.25}$$

This modification is not as simple as it first seems since the error is no longer convex with respect to μ_{m+1} and the minimization problem must be restricted. In our case, it empirically seems that imposing $|\mu_{m+1}| < 1$ is sufficient for the following estimator to be unique. We will suppose this. The estimator is then :

$$\hat{\mu}_{m+1} = \arg \min_{|\mu_{m+1}|<1} \frac{1}{N} \sum_{i=1}^{N} \sum_{n=m+1}^{d} -\log\left(|f_{i,m+1}(n)|^2 + |b_{i,m+1}(n)|^2\right). \tag{3.26}$$

Unfortunately, this estimator is not consistent and its asymptotic multiplicative bias should be corrected.

Proposition 3.22. *Let us define* $B_2 : a \mapsto \frac{2a\log(a)}{(1-a)^2} + \frac{1+a}{1-a}$ *and for* $y < 1$ *:*

$$g(y) = \frac{(1+y)^2 B_2^{-1}(y) - (1-y)^2}{(1+y)^2 B_2^{-1}(y) + (1-y)^2}$$

Then

$$\hat{\mu}_m \overset{a.s.}{\to} g^{-1}(|\mu_m|)\frac{\mu_m}{|\mu_m|}$$

Proof. See Appendix. □

The consistent version of (3.26) is then $\hat{\mu}_m^{(u)} = g(\hat{\mu}_m)\frac{\hat{\mu}_m}{|\hat{\mu}_m|}$.

Elliptical Energy

Let us finally introduce natural estimators for a specific dependent sample in the context of the elliptical distribution. We also prove that this estimator solves a minimum energy problem for an energy functional which we call elliptic energy.

As the forward and backward errors defined by equation 3.15 $e_m(n) := \begin{pmatrix} f_m(n) \\ b_m(n-1) \end{pmatrix}$

are 2-dimensional elliptical random vectors with known covariance (up to a multiplicative constant) given by equation 3.18 :

$$\mathbb{E}\left[e_m(n)e_m(n)^+\right] = \mathbb{E}[\tau^2]P_m \begin{pmatrix} 1 & -\mu_{m+1} \\ -\overline{\mu_{m+1}} & 1 \end{pmatrix}, \tag{3.27}$$

we can apply the elliptical approach given in [104] in order to estimate the covariance of this vector.
Let $y'_1, ..., y'_N$ be an iid sample following a centered elliptical distribution with scatter matrix given by equation (3.27). The likelihood of $z_1 = \frac{y'_1}{\|y'_1\|}, ..., \frac{y'_N}{\|y'_N\|}$ (whose distribution is complex ACG with scatter matrix proportional to 3.27) is written [105] :

$$l(z_1, ..., z_N) = \prod_{i=1}^{N} \frac{1-|\mu_{m+1}|^2}{2\pi^2} \left(z_i^+ \begin{pmatrix} 1 & \mu_{m+1} \\ \overline{\mu_{m+1}} & 1 \end{pmatrix} z_i \right)^{-2} \tag{3.28}$$

We remark that the likelihood is independent of τ and that $e_m(n)$ and $e_m(n')$ are dependent for $n \neq n'$. Define

$$\hat{\mu}_{m+1}^{(ell)} = \arg\min_{\mu_{m+1}\in\mathbb{C},|\mu_{m+1}|<1} \begin{array}{l} \sum_{n=m+2}^{d} \sum_{i=1}^{N} 2\log\left(e_{im}(n)^+ \begin{pmatrix} 1 & \mu_{m+1} \\ \overline{\mu_{m+1}} & 1 \end{pmatrix} e_{im}(n) \right) \\ -N(d-m-1)\log(1-|\mu_{m+1}|^2) \end{array} \tag{3.29}$$

This formula matches the maximum of the likelihood 3.28, would the $(e_{im}(n))_{i=1...N,n=m+2...d}$ be independent. Consistency of $\hat{\mu}_{m+1}^{(ell)}$ holds due to the independence of $e_{im}(n)$ and $e_{i'm}(n)$ for any n, m and $i \neq i'$.

Proposition 3.23. *The above estimator is consistent :*

$$\hat{\mu}_m^{(ell)} \overset{a.e.}{\to} \mu_m$$

Proof. Let note the 2×2 Hermitian Toeplitz matrix $\Sigma(\mu)$

$$\Sigma(\mu) = \begin{pmatrix} 1 & -\mu \\ -\overline{\mu} & 1 \end{pmatrix}.$$

Then

$$\Sigma(\mu)^{-1} = \frac{1}{1-|\mu|^2} \begin{pmatrix} 1 & \mu \\ \overline{\mu} & 1 \end{pmatrix}.$$

Moreover, we note $e_{in} = \begin{pmatrix} f_{i,m}(n) \\ b_{i,m}(n-1) \end{pmatrix}$ for $n = m+2...d$ and $i = 1...N$. $\hat{\mu}_m^{(ell)}$ is then the argument of the maximum of

$$l_N(\mu) = \frac{1}{N(d-m-1)} \sum_{n=m+2}^{d} \sum_{i=1}^{N} \log\left(\frac{e_{in}^+ \Sigma(\mu)^{-1} e_{in}}{|\Sigma^{-1}(\mu)|^{1/2} e_{in}^+ e_{in}} \right)$$

with respect to $\mu \in D = \{z \in \mathbb{C} \text{ s.t. } |z| < 1\}$. We will moreover note

$$l(\mu) = \mathbb{E}_Z \left[\log\left(\frac{Z^+\Sigma(\mu)^{-1}Z}{|\Sigma(\mu)^{-1}|^{1/2}} \right) \right]$$

where Z is an ACG-distributed random variable of scatter matrix $\Sigma(\mu)$.
We will prove that there exists a constant $C < 1$ such that for n large enough

$$|\hat{\mu}_N| < C \text{ a.s.}$$

Let us suppose the contrary. Then, there exists a subdivision φ and a collection of events Ω of non-zero measure such that $\forall \omega \in \Omega$

$$|\hat{\mu}_{\varphi(N)}(\omega)| \to 1.$$

In the following, we will omit the references to ω and φ.
Furthermore, the two eigenvalues of the matrix $\Sigma(\hat{\mu}_N)^{-1}$ are $\frac{1}{1-|\hat{\mu}_N|}$ and $\frac{1}{1+|\hat{\mu}_N|}$. Let us denote the projections of the respective eigenvectors on e_{in} by v_{in} and w_{in}.
Then

$$\begin{aligned} l_N(\hat{\mu}_N) &= \frac{1}{N(d-m-1)} \sum_{n=m+2}^{d} \sum_{i=1}^{N} \log\left((1-|\hat{\mu}_N|^2)^{1/2} \left(\frac{|v_{in}|^2}{1-|\hat{\mu}_N|} + \frac{|w_{in}|^2}{1+|\hat{\mu}_N|} \right) \right) \quad (3.30) \\ &\geq \frac{1}{N(d-m-1)} \sum_{n=m+2}^{d} \sum_{i=1}^{N} -\log\left(\frac{(1-|\hat{\mu}_N|}{1-|\hat{\mu}_N|^2)^{1/2}} \right) \\ &\geq \frac{1}{N(d-m-1)} \sum_{n=m+2}^{d} \sum_{i=1}^{N} -\log\left(\frac{(1-|\hat{\mu}_N|)^{1/2}}{(1+|\hat{\mu}_N|)^{1/2}} \right) \to +\infty. \end{aligned}$$

As $l_N(0) = 0$, this shows the contradiction. We can therefore restrict the minimization to the compact

$$M = \{ z \in \mathbb{C} \text{ s.t. } |z| < C \}.$$

Furthermore, we can note $\xi(x, \mu) = \log\left(\frac{x^+ \Sigma(\mu)^{-1} x}{|\Sigma^{-1}(\mu)|^{1/2} x^+ x} \right)$. Then

- $\mu \mapsto \xi(x, \mu)$ is continuous for all $x \in S_2 = \{ x \in \mathbb{C}^2 \text{ s.t. } \|x\| = 1 \}$
- $\xi(x, \mu) < \log((2/C)^{1/2} + 1)$ from the equation 3.30.

We can therefore apply the uniform law of large numbers to ξ on each subsamples $e_{1n}, ..., e_{Nn}$ for a fixed n since M is compact. This shows that

$$\begin{aligned} \sup_{\mu \in M} |l_N(\mu) - l(\mu)| &= \sup_{\mu \in M} \left| \frac{1}{N(d-m-1)} \sum_{i,n} \xi(\mu, e_{in}) - \mathbb{E}_Z[(\mu, Z)] \right| \\ &\leq \frac{1}{d-m-1} \sum_{n=m+2}^{d} \sup_{\mu \in M} \left| \frac{1}{N} \sum_{i=1}^{N} \xi(\mu, e_{in}) - \mathbb{E}_Z[(\mu, Z)] \right| \to_{\text{a.s.}} 0 \end{aligned}$$

Finally, as l is the expression of a likelihood, it is clear that for any $\epsilon > 0$

$$\inf_{\mu \in M \text{ s.t. } |\mu - \mu_{m+1}| > \epsilon} l(\mu) = \min_{\mu \in M \text{ s.t. } |\mu - \mu_{m+1}| > \epsilon} l(\mu) > l(\mu_0)$$

We can then apply Theorem 5.7 on M-estimators proposed by Van der Vaart [106]. □

Remark 3.37. *The estimator $\hat{\mu}_{m+1}^{(ell)}$ defined by equation 3.29 can be written*

$$\hat{\mu}_{m+1}^{(ell)} = \arg \min_{\mu_{m+1} \in \mathbb{C}, |\mu_{m+1}| < 1} \hat{U^{(m)}}$$

with $\hat{U^{(m)}}$ the empirical version of the following energy

$$U^{(m)} = (m+1-d)\log(1-|\mu_{m+1}|^2) + 2\sum_{n=m+1}^{d}\log\left(\begin{pmatrix} f_m(n) \\ b_m(n-1) \end{pmatrix}^+ \begin{pmatrix} 1 & \mu_{m+1} \\ \overline{\mu_{m+1}} & 1 \end{pmatrix} \begin{pmatrix} f_m(n) \\ b_m(n-1) \end{pmatrix}\right). \quad (3.31)$$

This criterion is the adaptation of the Burg criterion 3.14 for non-Gaussian vectors since we perform the same approximation for the maximum likelihood but for normalized version of the error vectors $e_m(n)$.

3.3 Optimization on Riemannian manifolds

The estimators of the scatter matrix proposed in the sequel are solutions to an optimization problem of the form :

$$\hat{\Sigma} = \arg \min_{\Sigma \in \mathcal{S}_d^{++}(\mathbb{C})} C(\Sigma)$$

for a cost function C supposed to be continuously differentiable where $\mathcal{S}_d^{++}(\mathbb{C})$ denotes the space of the Hermitian positive definite matrices.
We propose an algorithm in order to compute efficiently this minimum, or at least, a local minimum without convex assumption on C.
A first idea can be to search the roots of the gradient equation :

$$\nabla C(\hat{\Sigma}) = 0$$

For example, if C is the cost function of the likelihood of the ACG sample $(x_1, .., x_N)$, this equation is

$$\hat{\Sigma} = \frac{d}{N}\sum_{i=1}^{N}\frac{x_i x_i^+}{x_i^+ \Sigma^{-1} x_i}$$

The form of this equation leads to the Fixed Point algorithm proposed by Tyler [104]

$$\hat{\Sigma}_{t+1} = \frac{d}{N}\sum_{i=1}^{N}\frac{x_i x_i^+}{x_i^+ \hat{\Sigma}_t^{-1} x_i}$$

The proof of convergence of this algorithm is specific to the above formulation and can not be easily adapted for the computation of the minimum of any cost C.
We propose here to perform a gradient descent on the manifold of Hermitian positive definite matrices $\mathcal{S}_d^{++}(\mathbb{C})$. As the manifold is not a vector space, simple constrained Euclidean gradient descents are generally avoided because of their slow convergence properties. It is actually natural to perform a gradient descent in a Riemannian (Hermitian for the complex analogue) framework built for the manifold ($\mathcal{S}_d^{++}(\mathbb{C})$ for example).

We will first introduce some notions of Riemannian manifolds necessary to perform optimization algorithms and present the steepest descent algorithm in a general Riemannian framework proposed by Absil et al. [1].
We will then specify the metrics used for the computation of the estimated scatter matrix $\hat{\Sigma}$ in $\mathcal{S}_d^{++}(\mathbb{C})$ for the non stationary case and then in the manifold of reflection coefficients $\mathbb{R}_+^* \times D^{M-1}$ for the stationary case (where $D = \{z \in \mathbb{C} \text{ s.t. } |z| < 1\}$ is the Poincaré disk).
Let us recall that the reflection parameters are defined by the Levinson algorithm 3.13 as a re-parametrization of a Toeplitz matrix Σ.

Remark 3.38. *Actually, the scatter matrix Σ of a CES distribution is Hermitian positive definite with a trace condition $Tr(\Sigma) = d$ (see 3.7). However, we will define optimization procedures in the unconstrained space $\mathcal{S}_d^{++}(\mathbb{C})$. Indeed, we will see that the cost functions we introduce are independent with respect to a multiplication of the parameter Σ by a positive scalar. We can therefore choose a posteriori this multiplicative constant.*

3.3.1 Riemannian metrics, geodesics and exponential map

We will give a quick intuitive overview of some tools in Riemannian manifolds. For the sake of brievity, the precise definition of all introduced objects are not detailed here ; see [46] for details.

Metrics and tangent space

A Riemannian (respectively Hermitian) manifold V is a real (resp. complex) manifold equipped with a local inner product g_x on the tangent space T_xV defined at each point $x \in V$. We will assume for simplicity that V is an open submanifold of a vector space E of dimension d.
The tangent space is a vector space that contains the possible "directions" at which one can tangentially pass through x.

Definition 3.15. *Let $x \in V$. Then a vector $v \in E$ is a tangent vector to V if there exists a C^1 curve $\gamma :]-\epsilon;\epsilon[\to V$ such that $\gamma(0) = x$ and $\gamma'(0) = v$.*
The tangent space T_xV is the space of all tangent vectors v.

Remark 3.39. *Let $x \in V$, $v \in E$. As V is open in E, there exists ϵ such that $x + tv \in V$ for all $t \in]-\epsilon;\epsilon[$. Then if we define $\gamma(t) = x + tv$, $\gamma(0) = x$ and $\gamma'(0) = v$ i.e. $v \in T_xV$. Thus, if V is open in E, $T_xV = E$.*
If $C : V \to \mathbb{R}$ is a diffeomorphism, its differential in x d_xC is mapping from T_xV into $\mathbb{R}$ with for $v \in T_xV$

$$d_xC(v) = \lim_{\epsilon \to 0} \frac{C(x + \epsilon v) - C(x)}{\epsilon} = \langle \nabla C(x), v \rangle$$

with $\langle ., . \rangle$ the canonical scalar product of E.
For the open subset we will consider, this definition of a differentiate is sufficient. A general definition in a general context is available in [46].

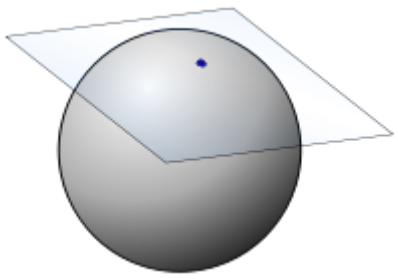


Figure 3.7 – Tangent space for a given point on a sphere

The inner product is called a Riemannian metric.

Definition 3.16. *A Riemannian metric g is the collection of g_x for all $x \in V$ such that g_x is a positive definite quadratic form on T_xV depending smoothly on x. Hence, g_x is an application $g_x : T_xV \times T_xV \to \mathbb{R}$ which satisfies :*

- g_x *is symmetric i.e. for any* $v, w \in T_xV$

$$g_x(v,w) = g_x(w,v)$$

- g_x *is bilinear i.e. for any* $a, b \in \mathbb{R}$ *and* $u, v, w \in T_xV$

$$g_x(av + bw, u) = ag_x(v,u) + bg_x(w,u)$$

- g_x *is positive definite i.e. for any* $v \in T_xV \backslash \{0\}$, $g_x(v,v) > 0$

Example 3.28. *Let* U *be an open subset of* $\mathbb{R}^d$ *and* $T_xU = \mathbb{R}^d$ *for all* $x \in U$. *If* $e_1, .., e_d$ *denotes the canonical basis of* $\mathbb{R}^d$, *we can define :*

$$g_x : T_xU \times T_xU \to \mathbb{R}, \left(\sum_{i=1}^d a_i e_i, \sum_{i=1}^d b_i e_i\right) \mapsto \sum_{i=1}^d a_i b_i$$

g_x *is a Riemannian metric. If* $U = \mathbb{R}^d$, *it is the canonical Euclidean metric.*

Example 3.29. *Let* $V = \mathbb{R} \times \mathbb{R}_+^*$ *and* $T_xV = \mathbb{R}^2$ *for all* $x \in V$. *we can define the metric in* $x = (\mu, \sigma) \in V$

$$g_x : T_xV \times T_xV \to \mathbb{R}, ((h_\mu, h_\sigma), (k_\mu, k_\sigma)) \mapsto \frac{h_\mu k_\mu}{\sigma} + \frac{h_\sigma k_\sigma}{2\sigma^2}$$

g_x *is a Riemannian metric. As we will see later, it corresponds to the Fisher metric associated to the univariate Gaussian model parametrized by a mean* μ *and a variance* σ.

The metric induces a norm on the tangent space at $x \in V$, T_xV

$$\|v\|_x = \sqrt{g_x(v,v)} \text{ for all } v \in T_xV$$

Remark that $\|v\|_x^2$ is often denoted by ds^2 in the Riemannian literature.

Exponential map and geodesics

Definition 3.17. *With a metric* g_x, *we define the length of a continuously differentiable path* $\gamma : [0;1] \to V$ *by*

$$L(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))} dt$$

with $\dot{\gamma}(t) = \lim_{\epsilon \to 0} \frac{\gamma(t+\epsilon) - \gamma(t)}{\epsilon} \in T_{\gamma(t)}V$.
The distance between x *and* $y \in V$ *is defined as the infimum of the length of all paths such that* $\gamma(0) = x$ *and* $\gamma(1) = y$. *The minimizing paths are called geodesics.*

If we are given an initial point $x \in V$ and a speed vector $v \in T_xV$, we suppose that there exists a unique geodesic γ_v satisfying

$$\begin{cases} \gamma(0) = x \\ \dot{\gamma}(0) = v \end{cases}.$$

Definition 3.18. *The final point of this "geodesic shooting" is referenced by the exponential map*

$$\exp_x(v) = \gamma(1)$$

The exponential map in $x \in V$ *is then an application* $\exp_x : T_xV \to V$

Remark 3.40. *The uniqueness of the geodesic allows us to define the inverse of an exponential map* $\exp^{-1}$ *also denoted by a* log *operator.*

$$\log_x = \exp_x^{-1} : V \to T_xV$$

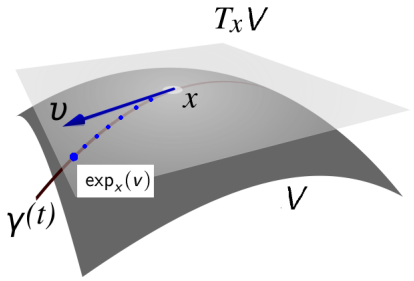


Figure 3.8 – Illustration of an exponential map defined on the tangent space of a manifold V

Operation	Euclidean geometry	Riemannian geometry
Shooting from x with speed $\overrightarrow{v}$	$x+\overrightarrow{v}$	$\exp_x(\overrightarrow{v})$
Speed vector shooting x into y	$\overrightarrow{v}=\overrightarrow{xy}$	$\overrightarrow{v}=\exp_x^{-1}(y)$
Distance between x and y	$\|x-y\|$	$\|\exp_x^{-1}(y)\|_x$

Table 3.2 – Link between notions in Riemannian and Euclidean geometry

Complex analogue of Riemannian tools

Hermitian metric

Until now, we have considered real Riemannian manifolds. However, for signal processing applications, the considered manifolds are complex ones. The complex analogue of Riemannian manifolds are called Hermitian manifolds. Let V be a Hermitian submanifold of a complex vector space E.
We can introduce in the same way a Hermitian metric that is a symmetrical sesquilinear positive definite form i.e. the symmetry of a metric $g_x : T_xV \times T_xV \to \mathbb{C}$ in $x \in V$ is replaced by :

$$g_x(v,w) = \overline{g_x(v,v)}$$

and g_x is positive definite if and only if $g_x(v,\overline{v}) > 0$ for all $v \in T_xV$.

Complex derivation

We have to introduce a natural notion of complex gradient. We give the method for a general cost function $C : \mu \in \mathbb{C} \mapsto C(\mu) \in \mathbb{R}$. We have no reason to consider C as an holomorphic function i.e. for $z, z_0 \in \mathbb{C}$

$$\frac{C(z) - C(z_0)}{z - z_0}$$

does not necessarily have a limit when $z \to z_0$. The idea is then naturally to exploit the bijection between $\mathbb{R}^2$ and $\mathbb{C}$. If we note $z = x + iy$, the hypothesis that C is differentiable almost everywhere in (x, y) is acceptable. The derivation of C with respect to x and y is called the $\mathbb{R}$-derivation in contrast to the $\mathbb{C}$-derivation that corresponds to the holomorphic derivation.
In order to carry tractable expression for the Taylor development of C, we introduce some notations. We can consider the cost function as a function of the couple $(z, \bar{z})$.

This consideration underlies the CR-calculus, also named Wirtinger calculus [73]. The $\mathbb{R}$-derivative and the conjugate $\mathbb{R}$-derivative of C are respectively given by :

$$\partial_z C = \left.\frac{\partial C}{\partial z}\right|_{\bar{z}=const} \text{ and } \partial_{\bar{z}} C = \left.\frac{\partial C}{\partial \bar{z}}\right|_{z=const}$$

These can be equivalently defined by

$$\partial_z C = \frac{1}{2}\left(\frac{\partial C}{\partial x} - i\frac{\partial C}{\partial y}\right) \text{ and } \partial_{\bar{z}} C = \frac{1}{2}\left(\frac{\partial C}{\partial x} + i\frac{\partial C}{\partial y}\right)$$

Moreover, we will exploit the fact that the costs expressed by C are real. This implies

$$\begin{cases} \overline{\partial_z C} = \partial_z \overline{C} = \partial_z C \\ \overline{\partial_{\bar{z}} C} = \partial_{\bar{z}} \overline{C} = \partial_{\bar{z}} C \end{cases}$$

We can now produce the Taylor expansion in terms of $\mathbb{R}$-derivative. Let $z = x + iy \in \mathbb{C}$ and $dz = dx + idy \in \mathbb{C}$ be a small increment. Then

$$\begin{aligned} C(z+dz) &= C(z) + \frac{\partial C}{\partial x} d_x + \frac{\partial C}{\partial y} d_y + o(\|dz\|) \\ &= C(z) + \Re\left[\left(\frac{\partial C}{\partial x} - i\frac{\partial C}{\partial y}\right)(d_x + i d_y)\right] + o(\|dz\|) \\ &= C(z) + 2\Re\left[(\partial_z C) dz\right] + o(\|dz\|) \\ &= C(z) + 2\Re\left[\overline{\partial_{\bar{z}} C} dz\right] + o(\|dz\|) \end{aligned}$$

This scalar product suggests to define as complex gradient of C

$$\nabla C = 2\partial_{\bar{z}} C.$$

In the following, we will omit the factor 2 that can be absorbed by the descent step. For Taylor expansions of higher order, we refer to Kreutz-Delgado [73]. Let us present nevertheless the gradient descent of order 2 for an Euclidean metric.

Euclidean complex gradient descent

If we consider an Euclidean metric, the first order gradient descent is then given by the "linear" approximation of the function C around the minimum $\hat{\mu}$. We initialize the algorithm with μ_0 and update μ_t (for $t > 0$) into :

$$\mu_{t+1} = \mu_t - \alpha \partial_{\bar{z}} C(\mu_t) \tag{3.32}$$

with $\alpha \in \mathbb{R}$ the step of descent.
Furthermore, we give the second order gradient descent :

$$\mu_{t+1} = \mu_t + \alpha \left(|\partial_{\bar{z}z} C|^2 - |\partial_{zz} C|^2\right)^{-1} \left(\partial_{\bar{z}\bar{z}} l \partial_z C - \partial_{\bar{z}z} C \partial_{\bar{z}} C\right) \tag{3.33}$$

with partial derivatives evaluated in μ_t :

$$\partial_{\bar{z}z} C = \frac{\partial}{\partial \bar{z}}\frac{\partial C}{\partial z}(\mu_t)\,;\, \partial_{zz} C = \frac{\partial}{\partial z}\frac{\partial C}{\partial z}(\mu_t)\,;\, \partial_{\bar{z}\bar{z}} C = \frac{\partial}{\partial \bar{z}}\frac{\partial C}{\partial \bar{z}}(\mu_t)$$

3.3.2 Minimization of a function on a Riemannian manifold : Riemannian steepest descent

Choice of the minimization algorithm

The minimization of a cost function on a Riemannian manifold was studied especially for the computation of geometric means. Indeed, if $\theta_1, ..., \theta_N$ are points of a Riemannian manifold Θ (for example Hermitian definite positive matrices), it is a natural question to search the mean of these points. A classical answer is the computation of the minimum of the L_2 distance [70]

$$\theta_{mean} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{N} d(\theta_i, \theta)^2 \tag{3.34}$$

where d is the Riemannian distance associated to Θ. If we replace the above minimization by

$$\theta_{median} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{N} d(\theta_i, \theta),$$

the solution is the geometric median introduced by Fréchet in his famous article [51]. The discrete mean 3.34 is a particular case of the definition of a general mean barycenter over a mass distribution μ defined on the parameter space Θ

$$\theta_{mean} = \arg\min_{\theta\in\Theta} \int_{\theta'\in\Theta} d(\theta, \theta')^2 d\mu(\theta').$$

The existence and uniqueness of such a minimum was proved by Cartan [36] if Θ is a complete, simply connected Riemannian manifold with non-positive curvature (so-called Cartan-Hadamard manifolds). The computation of this minimum can be done through a so-called Karcher flow [69]. The computation of the median equivalent of 3.3.2 by an adaptation of the Karcher flow was proposed in [111] ; see also [6] for an application in $\mathcal{S}_d^{++}(\mathbb{C})$. Mention furthermore stochastic versions of the Karcher flow proposed in [7] which was shown to converge on the circle (i.e. a Riemannian manifold with positive curvature).

The proof of convergence of the Karcher flow has not been achieved for an arbitrary cost function C. We then consider Riemannian line-search methods proposed by Absil et al. [1]. Furthermore, for constrained Hermitian definite positive matrices (for example if we add the Toeplitz structure), this allows us to perform a steepest descent preserving the structure [1] [22] [80].

Remark 3.41. *The steepest descent proposed by Absil et al. has been chosen for the simplicity of its implementation. However, let us mention the so-called Information Geometric Optimization (IGO) and the Geodesic Information Geometric Optimization (GIGO) proposed by Ollivier et al. [84] and Bensadon [17]. These promising optimization algorithms propose a general framework for stochastic optimization problems generalizing many existing stochastic optimization approaches. In particular, these algorithms aim to search the global optimum and not only a local optimum as it is the case for line-search algorithms.*

Presentation of the steepest descent

Let us consider a general cost function C defined on a Riemannian manifold V. We suppose that C is differentiable almost everywhere. Our aim is to find a critical point of C.

Definition 3.19. *A point x is called critical if $\nabla C(x) = 0$.*

Classical optimization algorithm often have their equivalent in a Riemannian framework. We propose to use a steepest descent. This allows us to compute a critical point of the function C ie a point x such that $\nabla C(x) = 0$. Note that we do not suppose that C is convex. The algorithm then computes local minima.

The algorithm performs an iterative geodesic shooting :

$$\Sigma_{t+1} = \exp_{\Sigma_t}(-\alpha_t \nabla C(\Sigma_t)) \tag{3.35}$$

with α_t the step of the descent computed for each $t > 0$, namely α_t should satisfy :

$$C(\exp_{\Sigma_t}(-\alpha_t \nabla C(\Sigma_t))) \leq C(\Sigma_t) - c\alpha_t \left\|\nabla C(\Sigma_t)\right\|_{\Sigma_t} \tag{3.36}$$

with $0 < c < 1$. This step size α_t is called Armijo step size. It allows to ensure that α_t suits in order to sufficiently decrease the cost function C. A classical way to search an Armijo step size is to successively divide α_t by 2 until the condition 3.36 is attained. This condition will be sufficient in order to have the convergence of the steepest descent algorithm.

Algorithm 2 Steepest descent for an arbitrary metric

Aim : Find a local minimum of C with a precision ϵ
Input : Σ_0 Initial state for $t = 0$, α_{max} maximum step size
while $\|\nabla C(\Sigma_t)\| > \epsilon$
  $\alpha_t = \alpha_{max}$
    while $C(\exp_{\Sigma_t}(-\alpha_t \nabla C(\Sigma_t))) > C(\Sigma_t) - c\alpha_t \left\|\nabla C(\Sigma_t)\right\|_{\Sigma_t}$
      $\alpha_t = \alpha_t/2$
    end
  $\Sigma_{t+1} = \exp_{\Sigma_t}(-\alpha_t \nabla C(\Sigma_t))$
  $t = t + 1$
end

We can in particular propose an alternative algorithm for the Fixed Point algorithm if we take the opposite of log-likelihood as function f to minimize :

$$\begin{cases} C(\Sigma) = d\sum_{i=1}^{N} \log(x_i^+ \Sigma^{-1} x_i) + N \log\det(\Sigma) \\ \nabla C(\Sigma) = \Sigma - \frac{d}{N}\sum_{i=1}^{N} \frac{x_i x_i^+}{x_i^+ \Sigma^{-1} x_i} \end{cases}$$

Standard convergence results of a gradient descent are available for the Riemannian framework.

Proposition 3.24. *Let $(\Sigma_t)_{t\geq 0}$ a sequence generated by Algorithm 2. Then every accumulation point of (Σ_t) is a critical point of f.*
If furthermore the level sets $\{x \text{ s.t. } C(x) \leq C(x_0)\}$ are compact then

$$\nabla C(\Sigma_t) \underset{t\to+\infty}{\longrightarrow} 0$$

Proof. see Theorem 4.3.1 and Corollary 4.3.2 of [1] □

3.3.3 Steepest descent for the manifold of Hermitian positive definite matrices

Once we consider a manifold, the question of the choice of the metric arises. We present two considerations leading to the same metric in the space of Hermitian positive definite matrices and compute the explicit formulation of the exponential map and the geodesics associated to this metric.

A metric invariant by group action

Our metric can satisfy a natural property, namely the invariance by affine change : if $z' = Az$ with $A \in GL_d(\mathbb{C})$ then $\mathbb{E}[z'z'^T] = A\mathbb{E}[zz^T]A^T = \mathbb{E}[zz^T]$.
This invariance corresponds to the invariance by the action of the group $GL_d(\mathbb{C})$ on $\mathcal{S}_d^{++}(\mathbb{C})$ defined by $A \star \Sigma = A\Sigma A^T$ for $A \in GL_d(\mathbb{C})$ and $\Sigma \in \mathcal{S}_d^{++}(\mathbb{C})$.
We then search a distance such that :

$$dist(\Sigma_1, \Sigma_2) = dist(\Sigma_1 \star A, \Sigma_2 \star A) \text{ for all } A \in GL_d(\mathbb{C})$$

According to Pennec et al.[90], this naturally leads to a unique metric.
By taking $A = \Sigma_1^{-1/2}$, we obtain a first simplification :

$$dist(\Sigma_1, \Sigma_2) = dist(Id, \Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}) := N(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})$$

Secondly, as unitary matrices verify $UU^+ = Id$, it holds $N(URU^+) = N(R)$ for all $U \in U_d$ (manifold of unitary matrices) by invariance of the distance by action of the group $GL_d(\mathbb{C})$. By using the spectral decomposition, we get :

$$dist(\Sigma_1, \Sigma_2) = N(\sigma_1, ..., \sigma_d) \text{ with } \sigma_1, ...\sigma_d \text{ the eigenvalues of } \Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}.$$

As $N(\sigma_1, ..., \sigma_d) = N(1/\sigma_1, ..., 1/\sigma_d)$, we naturally take :

$$dist(\Sigma_1, \Sigma_2) = (\sum_{i=1}^{n} log(\sigma_i)^2)^{1/2} = \|log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})\|_F \tag{3.37}$$

with $\|.\|_F$the classical Frobenius norm.

Remark 3.42. *This distance is also proposed in the book of Bhatia [21]. This is a particular case of a distance introduced by Siegel in the framework of symplectic geometry ; see [11] for details on this generalization.*

A metric coming from the information geometry

Amari [3] proposes to define a Riemannian metric from the Taylor expansion of a divergence function for a parametric family of distributions. Let us consider a general differentiable divergence D and a family of distributions parametrized by $\theta \in \Theta$ where Θ is included in a vector space of finite dimension. Then the Taylor expansion

$$D(P_\theta, P_{\theta+d\theta}) = d\theta^T G(\theta) d\theta$$

is a positive definite quadratic form with

$$G(\theta) = (g_{ij})_{i,j} = \left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} D(\theta, \eta)\Big|_{\eta=\theta} \right)_{i,j} \tag{3.38}$$

If D is the Kullback-Leibler divergence :

$$D(P,Q) = \int \log\left(\frac{dP}{dQ}\right) dP$$

Then, the resulting metric is called the Fisher metric because the matrix G is the Fisher information matrix associated to the family $(P_\theta)_{\theta\in\Theta}$

$$I(\theta) = \left(\int \frac{\partial \log p_\theta}{\partial\theta_i}(x)\frac{\partial \log p_\theta}{\partial\theta_j}(x)p_\theta(x)\right)_{i,j}$$

The resulting geometry is called the information geometry. If we compute the Fisher information matrix for multivariate centered Gaussian parametrized by their covariance Σ, the resulting norm on the tangent space is

$$\|d\Sigma\|_\Sigma^2 = Tr((\Sigma^{-1}d\Sigma)^2) \tag{3.39}$$

This norm induces the distance given above (see for example [12]). This metric can be associated to its dual metric when the Fisher matrix is replaced by the Hessian of the entropy [13]. Indeed, if we consider the Legendre dual of the entropy of a centered multivariate Gaussian distribution, its Hessian matrix is the Fisher information matrix.
We could argue that the Gaussian metric is not adapted for elliptical distributions. The advantage of Gaussian distributions is the simplicity of their metric. Let us nevertheless cite the existence of an explicit form of the information metric for elliptical distributions [18] of the form

$$\|d\Sigma\|_\Sigma^2 = \frac{1}{|H|^{2d}}\left(3b_h - \frac{1}{4}\right)\sum_{k=1}^{d}\frac{(d\lambda_k)^2}{\lambda_k^2} + 2\left(b_h - \frac{1}{4}\right)\sum_{k<l}\frac{d\lambda_k\lambda_l}{\lambda_k\lambda_l}$$

where the λ_k's are the eigenvalues of Σ and H be such that $H^+\Sigma H = \Lambda$, the diagonal matrix composed by the λ_k's ; see also [20] who proposed another approach for the computation of the information matrix of these distributions.

Geodesics and exponential map in $\mathcal{S}_d^{++}(\mathbb{C})$
With the distance defined above, the geodesics and also the exponential map are explicitly known.

For any point of the manifold, the tangent space to $\mathcal{S}_d^{++}(\mathbb{C})$ is the space of Hermitian matrices $\mathcal{S}_d(\mathbb{C})$. Let Σ be a path with values in $\mathcal{S}_d^{++}(\mathbb{C})$. We get the following geodesics equations :

$$\begin{cases} \frac{d}{dt}(\Sigma^{-1}\dot{\Sigma}) = 0 \\ Tr((\Sigma^{-1}\dot{\Sigma})^2) = 0 \end{cases}$$

These equations are explicitly resolved, and if $\Sigma(0) = \Sigma_1$ and $\Sigma(1) = \Sigma_2$:

$$\Sigma(t) = \Sigma_1^{1/2}(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})^t\Sigma_1^{1/2} \text{ for } t \in [0,1]$$

It is interesting to note that the "middle" of this geodesic is $(\Sigma_1^{1/2}(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})^{1/2}\Sigma_1^{1/2})$ and corresponds to a geometric symmetrized mean. If Σ_1 and Σ_2 could commute, we would get $(\Sigma_1\Sigma_2)^{1/2}$.

By deriving this path in $t = 0$, we get :

$$\dot{\Sigma}(0) = W = \Sigma_1^{1/2} log((\Sigma_1^{-1/2})^+\Sigma_2\Sigma_1^{-1/2})R_1^{1/2}$$

We deduce that if we shoot from Σ with the speed $\dot{\Sigma}(0) = W$ which is a matrix in the tangent space, we get the exponential map (see Definition 3.18) :

$$\exp_\Sigma(W) = \Sigma(1) = \Sigma^{1/2} \exp(\Sigma^{-1/2} W \Sigma^{-1/2}) \Sigma^{1/2}.$$

The inverse exponential map is given by :

$$\exp_\Sigma^{-1}(S) = \log_\Sigma(S) = \Sigma^{1/2} \log(\Sigma^{-1/2} S \Sigma^{-1/2}) \Sigma^{1/2}. \tag{3.40}$$

This exponential map gives a correspondence between the tangent space and $\mathcal{S}_d^{++}(\mathbb{C})$. This allows us to perform geodesic shooting. This is the main tool of the Hermitian gradient descent.

Algorithm 3 Steepest descent for the metric defined by the equation 3.39

Aim : Find a local minimum of $C : \mathcal{S}_d^{++}(\mathbb{C}) \to \mathbb{R}$ with a precision ϵ
Input : Σ_0 Initial state for $t = 0$, α_{max} maximum step size
while $\|\nabla C(\Sigma_t)\| > \epsilon$
  $\alpha_t = \alpha_{max}$
    while $C(\Sigma_t^{1/2} e^{-\alpha_t \Sigma_t^{-1/2} \nabla C(\Sigma_t)) \Sigma_t^{-1/2}} \Sigma_t^{1/2}) > C(\Sigma_t) - c\alpha_t Tr\left[\left(\Sigma_t^{-1} \nabla C(\Sigma_t)\right)^2\right]$
      $\alpha_t = \alpha_t / 2$
    end
  $\Sigma_{t+1} = \Sigma_t^{1/2} e^{-\alpha_t \Sigma_t^{-1/2} \nabla C(\Sigma_t)) \Sigma_t^{-1/2}} \Sigma_t^{1/2}$
  $t = t + 1$
end

3.3.4 Steepest descent in the Poincaré disk

In practice, performing an Euclidean steepest descent in the unit disk

$$D = \{\mu \in \mathbb{C} \text{ s.t. } |\mu| < 1\}$$

does not guarantee a quick convergence, especially if the solution has a modulus close to one. We will take into account the natural geometry of the parameters of the autoregressive model. Indeed, a natural information geometry can be associated to any parametric model by defining a Riemannian metric through the Hessian of the entropy function (see [6] [12]).Note that this geometry is different to the Fisher information geometry even if closely related.
Fortunately, the geometry associated to the autoregressive model reparametrized by reflection parameters is simple in the sense that the geodesics does not have cross products and, for each reflection parameter μ_k, the resultant metric is the Poincaré metric in the unit disc D :

$$d\mu^2 = \frac{|d\mu|^2}{(1 - |\mu|^2)^2} \tag{3.41}$$

Yang [111] gives an explicit formula of the geodesics associated to this metric :

$$\exp_\mu(v) = \frac{(\mu + e^{i\theta}) e^{2|v|_D} + (\mu - e^{i\theta})}{(1 + \overline{\mu} e^{i\theta}) e^{2|v|_D} + (1 - \overline{\mu} e^{i\theta})} \tag{3.42}$$

where $\theta = \arg(v)$ and $|v|_D = \frac{|v|}{1-|\mu|^2}$.
Let $C : D \to \mathbb{R}$ be a cost function to minimize. With the above exponential map, the

complex Riemannian steepest descent is then performed by Algorithm 4. The specifications of this algorithm for different cost functions C are done in the next Section 3.4.

Algorithm 4 Steepest descent in the Poincare disk for the metric defined by equation 3.41

Aim : Find a local minimum of $C : D \to \mathbb{R}$ with a precision ϵ

Input : μ_0 Initial state for $t = 0$, α_{max} maximum step size, the exponential map is given by equation 3.42

while $\|\nabla C(\mu_t)\| > \epsilon$

$\alpha_t = \alpha_{max}$

while $C(\exp_{\mu_t}(-\alpha_t \nabla C(\mu_t))) > C(\mu_t) - c\alpha_t \|\nabla C(\mu_t)\|_{\mu_t}$

$\alpha_t = \alpha_t/2$

end

$\mu_{t+1} = \exp_{\mu_t}(-\alpha_t \nabla C(\mu_t))$

$t = t + 1$

end

3.4 Summary of the Burg algorithms

We will sum up the computation of the three proposed algorithms

- Normalized Burg
- Log Burg
- Elliptical Burg

Each Burg algorithm is related to a recursive estimation of the reflection parameters appearing naturally in the Levinson algorithm :

Algorithm 5 Generalized Burg-Levinson algorithm

Aim : Estimation of the autoregressive parameters $(P_M, a_1^{(M)}, ..., a_M^{(M)})$ and, equivalently, of $(P_0, \mu_1, ..., \mu_M)$

Input : a sample of N vectors $(x_1, .., x_N)$ in $\mathbb{C}^d$, the order of the underlying autoregressive process M

$P_0 = \frac{1}{Nd} \sum_{i=1}^{N} \sum_{k=1}^{d} |x_{ik}|^2$

For $1 \leq i \leq N$ and $1 \leq n \leq d$, $\begin{cases} f_{i,0}(n) = x_{in} \\ b_{i,0}(n) = x_{in} \end{cases}$

for $m = 1...M$

Estimation of $\hat{\mu}_m$ from f_{m-1} and b_{m-1} (through Normalized, Log or Elliptic Burg estimator)

$P_m = (1 - |\hat{\mu}_m|^2) P_{m-1}$

$$\begin{pmatrix} a_1^{(m)} \\ \vdots \\ a_{m-1}^{(m)} \end{pmatrix} = \begin{pmatrix} a_1^{(m-1)} \\ \vdots \\ a_{m-1}^{(m-1)} \end{pmatrix} + \hat{\mu}_m \begin{pmatrix} \overline{a}_{m-1}^{(m-1)} \\ \vdots \\ \overline{a}_1^{(m-1)} \end{pmatrix}$$

$a_m^{(m)} = \hat{\mu}_m$

Forward and backward errors for $1 \leq i \leq N$ and $m + 1 \leq n \leq d$,

$\begin{cases} f_{i,m}(n) = f_{i,m-1}(n) + \hat{\mu}_m b_{i,m-1}(n-1) \\ b_{i,m}(n) = b_{i,m-1}(n-1) + \overline{\hat{\mu}_m} f_{i,m-1}(n) \end{cases}$

end

The remaining point is then the estimation of the reflection parameters for each $1 \leq m \leq M$. Let us recall the exponential map given in equation 3.42

$$\exp_\mu(v) = \frac{(\mu + e^{i\theta})e^{2|v|_D} + (\mu - e^{i\theta})}{(1 + \overline{\mu}e^{i\theta})e^{2|v|_D} + (1 - \overline{\mu}e^{i\theta})}$$

where $\theta = \arg(v)$ and $|v|_D = \frac{|v|}{1-|\mu|^2}$.

Algorithm 6 Normalized Burg estimation

Aim : Estimation of the m-th coefficient of reflection $\hat{\mu}_{m+1}$
Input : forward and backward errors $f_{i,m}(n)$ and $b_{i,m}(n)$
$z = -\frac{2}{N(d-m-1)} \sum_{i=1}^{N} \sum_{n=m+2}^{d} \frac{\overline{b_{i,m}(n-1)} f_{i,m}(n)}{|f_{i,m}(n)|^2 + |b_{i,m}(n-1)|^2}$
$B_1 = x \mapsto \frac{1-x^2}{x}\left(\frac{\log(1-x)-\log(1+x)}{2x} + \frac{1}{1-x^2}\right)$
$\hat{\mu}_{m+1} = B_1^{-1}(|z|)\frac{z}{|z|}$

Algorithm 7 Elliptic Burg estimation

Aim : Estimation of the m-th coefficient of reflection $\hat{\mu}_{m+1}$
Input : forward and backward errors $f_{i,m}(n)$ and $b_{i,m}(n)$, an initial state z_0, a descent step α, a tolerance η
while $|\delta_t| > \eta$
$c_i(n) = f_{i,m}(n)\overline{b_{i,m}}(n-1)$ for $m+2 \leq n \leq d$
$d_i(n) = |f_{i,m}(n)|^2 + |b_{i,m}(n-1)|^2$ for $m+2 \leq n \leq d$
$\delta_t = \frac{2}{N(d-m-1)} \sum_{n=m+2}^{d} \sum_{i=1}^{N} \frac{c_i(n)}{d_i(n) + 2\Re(\overline{z_t} c_i(n))} + \frac{z_t}{1-|z_t|^2}$
$z_{t+1} = \exp_{z_t}(-\alpha \delta_t)$
end
$\hat{\mu}_{m+1} = z_t$

Algorithm 8 Log Burg estimation

Aim : Estimation of the m-th coefficient of reflection $\hat{\mu}_{m+1}$
Input : forward and backward errors $f_{i,m}(n)$ and $b_{i,m}(n)$, an initial state z_0, a descent step α, a tolerance η
while $|\delta_t| > \eta$
$c_i(n) = f_{i,m}(n)\overline{b_{i,m}(n-1)}$ for $m+2 \leq n \leq d$
$d_i(n) = |f_{i,m}(n)|^2 + |b_{i,m}(n-1)|^2$ for $m+2 \leq n \leq d$
$\delta_t = \frac{1}{N(d-m-1)} \sum_{n=m+2}^{d} \sum_{i=1}^{N} \frac{2c_i(n) + z_t d_i(n)}{d_i(n) + 4\Re(\overline{z_t} c_i(n)) + |z_t|^2 d_i}$
$z_{t+1} = \exp_{z_t}(-\alpha \delta_t)$
end
$B_2 : a \mapsto \frac{2a\log(a)}{1-a)^2} + \frac{1+a}{1-a}$
$g : y \mapsto \frac{(1+y)^2 B_2^{-1}(y) - (1-y)^2}{(1+y)^2 B_2^{-1}(y) + (1-y)^2}$
$\hat{\mu}_{m+1} = g^{-1}(|z_t|)\frac{z_t}{|z_t|}$

Remark 3.43. *The step*

$$P_0 = \frac{1}{Nd} \sum_{i=1}^{N} \sum_{k=1}^{d} |x_{ik}|^2$$

of the Generalized Burg-Levinson algorithm should be replaced by

$$P_0 = median\left(\frac{1}{d}\sum_{k=1}^{d}|x_{ik}|^2\right)$$

for a gain of robustness.

3.5 Regularization for Burg estimators

The Burg iteration procedure has a major drawback. It cumulates the errors especially when the order of the autoregressive model is false and the number of samples, namely N, is small. In practice, we often overestimate this order and the estimated reflection parameters of higher order have large modulus even if they are not relevant for the signal. A classical method for limiting the impact of the reflection parameters of higher order is to impose a regularity constraint (or smoothness prior) on the spectrum. We will consider the following smoothing measure of the autoregressive model spectrum parametrized by $a_1^{(m)},\dots,a_m^{(m)}$ (see for example [10]) :

$$C_m = \int_{-1/2}^{1/2}|A_m(f)|^2 df = \sum_{k=0}^{m}|a_k^{(m)}|^2$$

with $A_m(f) = \sum_{k=0}^{m} a_k^{(m)} e^{-2i\pi kf}$.
In the estimation process, instead of minimizing

$$\hat{\mu}_m = \arg\min_{\mu_m\in D}\hat{U}^{(m)}(\mu_m)$$

we add a regularity term controlled by a parameter γ

$$\hat{\mu}_m = \arg\min_{\mu_m\in D}\hat{U}^{(m)}(\mu_m) + \gamma C_m(\mu_m)$$

Let recall that C_m depends on μ_m in the following manner :

$$C_m(\mu_m) = |\mu_m|^2 + 1 + \sum_{k=1}^{m-1}\left|a_k^{(m-1)} + \mu_m\overline{a_{m-k}^{(m-1)}}\right|^2$$

We will apply the regularization process to the Normalized Burg estimator and the Elliptical Burg estimator.

3.5.1 Regularized Gaussian Burg estimator

For the classical Burg error, the regularized estimator can be explicitly expressed by resolving the minimization process :

$$\hat{\mu}_m = \arg\min_{\mu_m}\frac{1}{N}\sum_{i=1}^{N}\sum_{n=m+1}^{d}|f_{i,m}(n)|^2 + |b_{i,m}(n)|^2 + \gamma\sum_{k=0}^{m}|a_k^{(m)}|^2$$

Proposition 3.25. *The Regularized Burg estimator is given by :*

$$\hat{\mu}_m = -\frac{\frac{2}{N}\sum_{i=1}^{N}\sum_{n=m+1}^{d}\overline{b_{i,m-1}(n-1)}f_{i,m-1}(n) + \gamma\sum_{k=1}^{m-1}a_k^{(m-1)}a_{m-k}^{(m-1)}}{\frac{1}{N}\sum_{i=1}^{N}\sum_{n=m+1}^{d}|b_{i,m-1}(n-1)|^2 + |f_{i,m-1}(n)|^2 + \gamma\sum_{k=0}^{m-1}|a_k^{(m-1)}|^2} \tag{3.43}$$

Proof. The solution of the minimization problem is the solution of the gradient equation :

$$\nabla_{\mu_m} U^{(m)}(\hat{\mu}_m) + \gamma \nabla_{\mu_m} C_m(\hat{\mu}_m) = 0$$

Furthermore

$$\begin{cases} \nabla_{\mu_m} \hat{U}^{(m)}(\hat{\mu}_m) = \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{n=m+1}^{d} |b_{i,m-1}(n-1)|^2 + |f_{i,m-1}(n)|^2 \right) \hat{\mu}_m \\ + \frac{2}{N} \sum_{i=1}^{N} \sum_{n=m+1}^{d} \overline{b_{i,m-1}(n-1)} f_{i,m-1}(n) \\ \nabla_{\mu_m} C_m(\hat{\mu}_m) = \hat{\mu}_m \sum_{k=0}^{m-1} |a_k^{(m-1)}|^2 + \sum_{k=1}^{m-1} a_k^{(m-1)} a_{m-k}^{(m-1)} \end{cases}$$

Therefore there is a unique solution to the gradient equation. □

Remark 3.44. *We remark that the estimator given by the equation 3.43 has a regularization parameter γ combined with a quadratic term. In practice, we choose $\gamma = \gamma_0 \frac{1}{N} \sum_{i=1}^{N} \|x_i\|^2$ with γ_0 a free parameter. This allows the regularization to be independent of the power of the signal which is often a requirement for applications in radar since it allows to preserve the Constant False Alarm Rate (CFAR) property of the estimator.*

3.5.2 Regularized Normalized Burg estimator

The Regularized Normalized Burg estimator is the solution of :

$$\hat{\mu}_m = \arg\min_{\mu_m} \frac{1}{N} \sum_{i=1}^{N} \sum_{n=m+1}^{d} \frac{|f_{i,m}(n)|^2 + |b_{i,m}(n)|^2}{|f_{i,m-1}(n)|^2 + |b_{i,m-1}(n-1)|^2} + \gamma \sum_{k=0}^{m} |a_k^{(m)}|^2 \tag{3.44}$$

Proposition 3.26. *The biased Regularized Normalized Burg estimator is given by :*

$$\hat{\mu}_m = - \frac{\frac{2}{N} \sum_{i=1}^{N} \sum_{n=m+1}^{d} \frac{\overline{b_{i,m-1}(n-1)} f_{i,m-1}(n)}{|f_{i,m-1}(n)|^2 + |b_{i,m-1}(n-1)|^2} + \gamma \sum_{k=1}^{m-1} a_k^{(m-1)} a_{m-k}^{(m-1)}}{(d-m) + \gamma \sum_{k=0}^{m-1} |a_k^{(m-1)}|^2} \tag{3.45}$$

If $\gamma = 0$, the following estimator is unbiased :

$$\hat{\mu}_m^{(u)} = B_1^{-1}(|\hat{\mu}_m|) \frac{\hat{\mu}_m}{|\hat{\mu}_m|} \tag{3.46}$$

with B_1 given by equation 3.22.

Proof. The solution of the minimization problem is the solution of the gradient equation :

$$\nabla_{\mu_m} U^{(m)}(\hat{\mu}_m) + \gamma \nabla_{\mu_m} C_m(\hat{\mu}_m) = 0$$

Furthermore

$$\begin{cases} \nabla_{\mu_m} \hat{U}^{(m)}(\hat{\mu}_m) = (d-m)\hat{\mu}_m + \frac{2}{N} \sum_{i=1}^{N} \sum_{n=m+1}^{d} \frac{\overline{b_{i,m-1}(n-1)} f_{i,m-1}(n)}{|f_{i,m-1}(n)|^2 + |b_{i,m-1}(n-1)|^2} \\ \nabla_{\mu_m} C_m(\hat{\mu}_m) = \hat{\mu}_m \sum_{k=0}^{m-1} |a_k^{(m-1)}|^2 + \sum_{k=1}^{m-1} a_k^{(m-1)} a_{m-k}^{(m-1)} \end{cases}$$

This concludes the first part of the proposition. The unbiasedness comes from the non regularized Normalized Burg estimator. □

3.5.3 Regularized Elliptical Burg estimator

With the same reasoning, we give the regularized version of the Elliptical Burg estimator.

Algorithm 9 Minimum of Regularized Elliptical Burg Error

Aim : Compute the minimum of Regularized Elliptical Burg Error $\hat{\mu}_m$
Input : an initial state z_0, a descent step α, a tolerance η
while $|\delta_t| > \eta$
$c_i(n) = f_{i,m-1}(n)\overline{b_{i,m-1}}(n-1)$ for $m+1 \leq n \leq d$
$d_i(n) = |f_{i,m-1}(n)|^2 + |b_{i,m-1}(n-1)|^2$ for $m+1 \leq n \leq d$
$\delta_t^{(U)} = \frac{2}{N(d-m)} \sum_{n=m+1}^{d} \sum_{i=1}^{N} \frac{c_i(n)}{d_i(n)+2\Re(\overline{z_t}c_i(n))} + \frac{z_t}{1-|z_t|^2}$
$\delta_t^{(C)} = z_t \sum_{k=0}^{m-1} |a_k^{(m-1)}|^2 + \sum_{k=1}^{m-1} a_k^{(m-1)} a_{m-k}^{(m-1)}$
$\delta_t = \delta_t^{(U)} + \gamma \delta_t^{(C)}$
$z_{t+1} = \exp_{z_t}(-\alpha \delta_t)$
end
$\hat{\mu}_m = z_t$

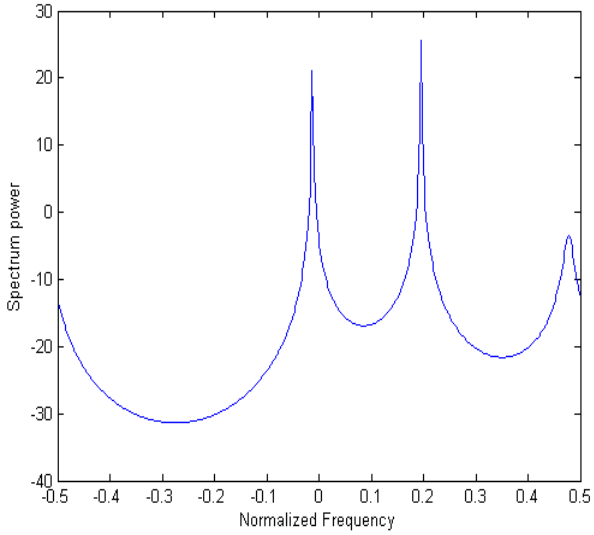
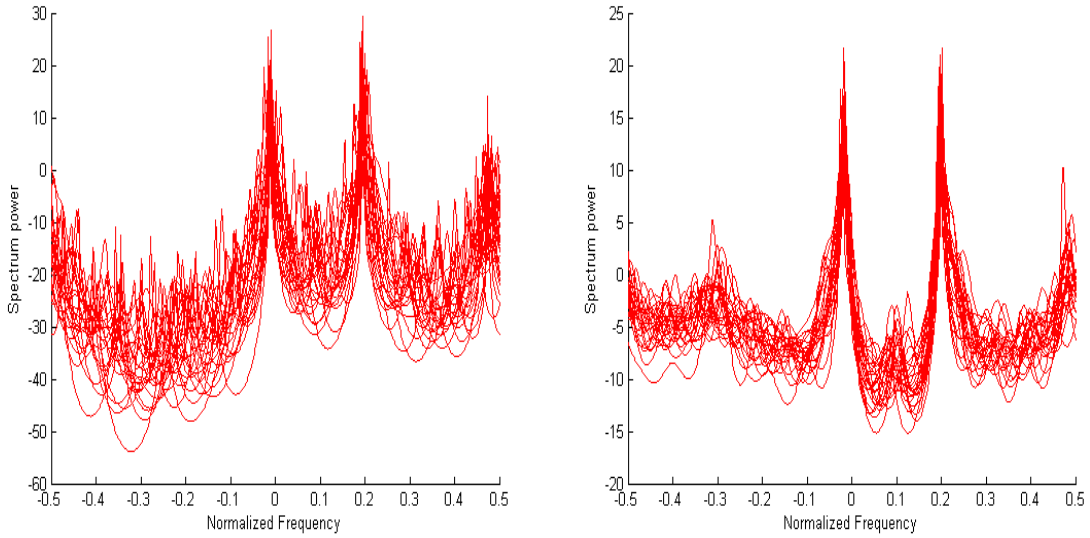


FIGURE 3.9 – Simulated autoregressive process spectra

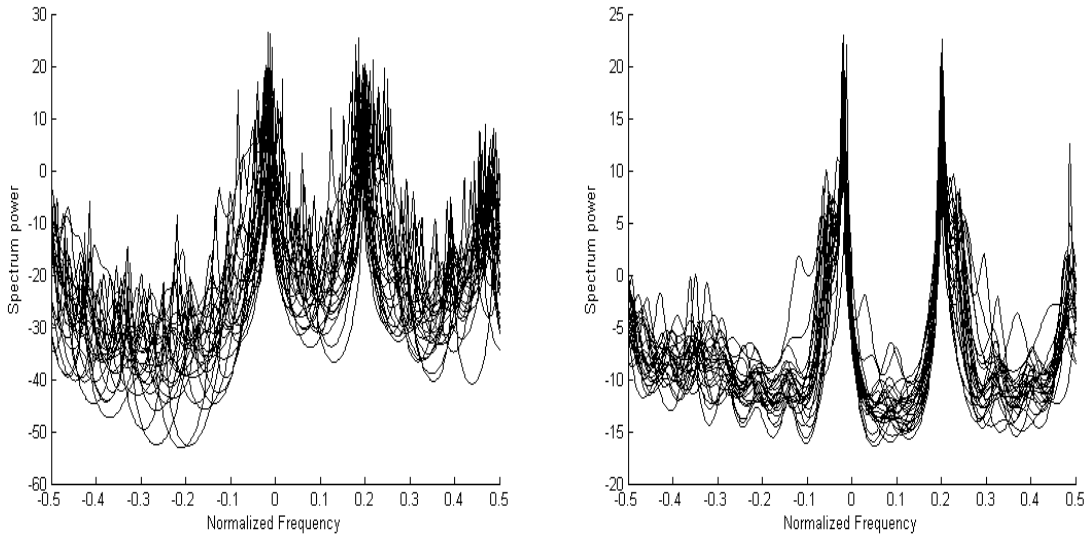
3.5.4 Illustration

In order to illustrate the effect of the regularization, we simulate an autoregressive vector of order 3 with $d = 15, P_0 = 1, \mu_1 = -0.6 - 0.5i, \mu_2 = 0.3 + 0.8i, \mu_3 = 0.5 + 0.8i$. We perform the estimation with $N = 4$ for 25 samples. As the order is unknown, we estimate an autoregressive vector of maximal order 14 which corresponds to estimate the covariance of a stationary vector.

In Figure 3.10, the regularization highlights the three simulated frequencies in the signal. However, the global level of the spectra is increased. The same drawback exists for the regularized version of the classical Burg algorithm. This is explained by the bias of the estimated reflection parameters in the regularized framework. The low standard deviations, especially for high order coefficients, compensate this bias effect (see Table 3.3). We observe a compromise bias/variance modulated by the regularization parameter γ.



(a) Regularized Burg Elliptical estimator with $\gamma = 0$ (b) Regularized Burg Elliptical estimator with $\gamma = 1$ (no regularization)



(c) Regularized Burg Normalized estimator with $\gamma = 0$(d) Regularized Burg Normalized estimator with $\gamma = 5$ (no regularization)

Figure 3.10 – Autoregressive spectra of Regularized Burg Elliptical estimator and Regularized Burg Normalized estimator

3.6 Elliptical models and robustness to heavy contamination

3.6.1 Two classes of robust M-estimators

In order to deal with the gross error model, we need a robust estimator. In the literature, the general class of M-estimators obtained as the minima of sums of the functions of the data is often considered (see for example Huber and Ronchetti [66]). More specifi-

Estimator	$\hat{\mu}_1/\sigma_{\hat{\mu}_1}$	$\hat{\mu}_2/\sigma_{\hat{\mu}_2}$	$\hat{\mu}_3/\sigma_{\hat{\mu}_3}$	$\hat{\mu}_{14}/\sigma_{\hat{\mu}_{14}}$
RBE with $\gamma = 0$	-0.60-0.51i/0.17	0.29+0.79i/0.10	0.49+0.77i/0.07	0.04+0.01i/0.45
RBE with $\gamma = 1$	-0.36-0.29i/0.13	0.05+0.05i/0.10	-0.03+0.47i/0.09	-0.01-0.01i/0.15
RBN with $\gamma = 0$	-0.58-0.52i/0.21	0.29+0.80i/0.11	0.49+0.73i/ 0.11	-0.11-0.02i/0.54
RBN with $\gamma = 5$	-0.48-0.38i / 0.25	0.15+0.35i/0.14	0.02+0.61i/0.09	-0.06+0.06i/0.10

TABLE 3.3 – Estimated mean and standard deviation of the reflection parameters for the Regularized Burg Elliptical estimator (RBE) and the Regularized Burg Normalized estimator (RBN)

cally, we are interested in M-estimators expressed as minima of the empirical version of a divergence between two measures. The reason for the popularity of these methods is the expression of the maximum likelihood as the minimization of the empirical version of the Kullback-Leibler divergence. It is then natural to generalize this approach by introducing other divergences.
Broniatowski and Vajda [30] listed four different kind of divergence criteria. We will detail and use two of them for their simplicity and their flexibility for the choice of the compromise between robustness and efficiency :

- a ψ-divergence inducing a non-normalized estimator for the scatter matrix Σ
- a γ-divergence inducing a normalized estimator for Σ

Non normalized M-estimator

Basu et al. [15] proposed to consider the minimization of a power divergence :

$$d_\alpha(f,g) = \int \left[f(z)^{1+\alpha} - \left(1 + \frac{1}{\alpha}\right) g(z) f(z)^\alpha + \frac{1}{\alpha} g(z)^{1+\alpha} \right] dz$$

Eguchi and Kano [48] proposed to consider a generalization of power divergences based on the log-likelihood

Definition 3.20. *Let ψ be a differentiable, strictly increasing and convex function, $\Psi^* : z \mapsto \int_0^z \exp(s)\psi'(s)ds$ and f and g two density functions with respect to the Lebesgue measure. We define a ψ-divergence by*

$$d_\psi(f,g) = \int \psi(\log g(z)) g(z) dz - \int \psi(\log f(z)) g(z) dz + \int \Psi^*(\log f(z)) dz - \int \Psi^*(\log g(z)) dz \tag{3.47}$$

The formulation of this divergence (in particular the role of Ψ^*) is explained by the following proposition.

Proposition 3.27. *d_ψ is a divergence in the sense that for the two densities f and g*

- $d_\psi(f,g) \geq 0$
- $d_\psi(f,g) = 0$ *if and only if* $f = g$

Proof. This is clear if we write the divergence :

$$d_\psi(f,g) = \int \delta_\psi(f(z), g(z)) dz$$

with for $t, u > 0$

$$\delta_\psi(t,u) = \int_t^u \frac{u-v}{v} \psi'(\log(v)) dv$$

□

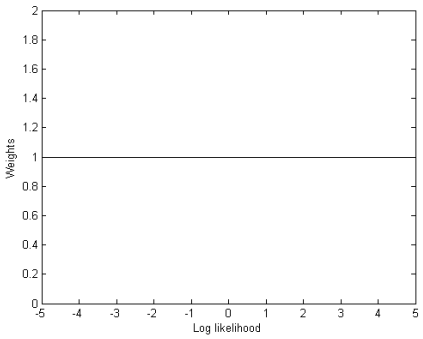


Figure 3.11 – Weighting function of the Kullback-Leibler divergence

The empirical version of 3.47 is easy to formulate. In the case of an elliptical distribution, the estimation of the parameter Σ_0 through the minimization of the divergence will be :

$$\hat{\Sigma}_0 = \arg\min_{\Sigma} -\frac{1}{N}\sum_{i=1}^{N}\psi(\log f_\Sigma(z_i)) + \int \Psi^*(\log f_\Sigma(z))dz \tag{3.48}$$

Eguchi and Kano showed that this estimator is Fisher-consistent [48], i.e. if we note for any distribution P

$$\hat{\Sigma}_0(P) := \arg\min_{\Sigma} -\int \psi(\log f_\Sigma(z))dP(z) + \int \Psi^*(\log f_\Sigma(z))dz,$$

it holds that $\hat{\Sigma}_0(F_{\Sigma_0}) = \Sigma_0$.
$\hat{\Sigma}_0$ also satisfies an estimating equation based on the score function $s(z,\Sigma) = \partial_\Sigma[\log(f_\Sigma)](z)$:

$$\frac{1}{N}\sum_{i=1}^{N}\psi'(\log f_{\hat{\Sigma}_0}(z_i))s(z_i,\hat{\Sigma}_0) = \mathbb{E}\left[\psi'(\log f_{\hat{\Sigma}_0}(z))s(z,\hat{\Sigma}_0)\right]$$

Remark 3.45. *This approach is considered as a M-estimation [66]. ψ' can be considered as a weighting function of the log-likelihood of the sample.*

Example 3.30. *(Kullback-Leibler divergence)*
If we choose $\psi(x) = x$, the estimation corresponds to the minimization of the Kullback-Leibler divergence i.e. the maximum likelihood estimator :

$$\hat{\Sigma}_0 = \arg\min_{\Sigma} -\frac{1}{N}\sum_{i=1}^{N}s(z_i,\Sigma) = \arg\max_{\Sigma}\frac{1}{N}\sum_{i=1}^{N}s(z_i,\Sigma)$$

Example 3.31. *(L_2-divergence)*
The L_2-divergence corresponds to the case $\psi = \exp$. The estimator satisfies then :

$$\hat{\Sigma}_0 = \arg\min_{\Sigma} -\frac{1}{N}\sum_{i=1}^{N}f_\Sigma(z_i) + \frac{1}{2}\int f_\Sigma(z)^2 dz$$

which is known to be robust but not efficient (see Hampel et al. [60])

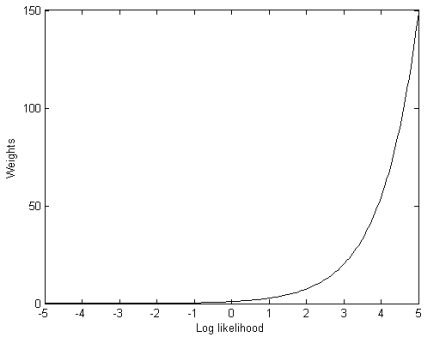


FIGURE 3.12 – Weighting function of the L_2 divergence

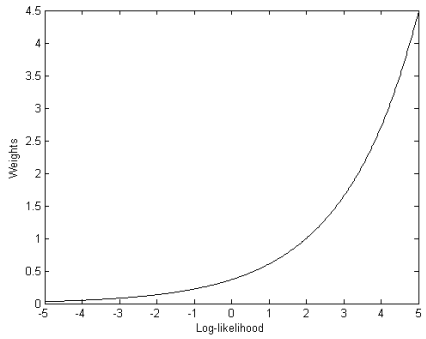


FIGURE 3.13 – Weighting function of the Basu divergence for $\beta = 0.5$ and $\mu = 2$

Example 3.32. *(Basu divergence)*
The Basu divergence corresponds to the case $\psi = x \mapsto \frac{1}{\beta}\exp(\beta(x-\mu))$. *The estimator then satisfies :*

$$\hat{\Sigma}_0 = \arg\min_{\Sigma} e^{-\beta\mu}\left[-\frac{1}{N\beta}\sum_{i=1}^{N} f_\Sigma(z_i)^\beta + \frac{1}{\beta+1}\int f_\Sigma(z)^{\beta+1}dz\right]$$

We add the parameter μ *for a relocation of the score. The more* μ *is high, the easier it will be to estimate the term* $\int \Psi^*(\log f_\Sigma(z))dz$ *by Monte Carlo.*

Example 3.33. *(Trimmed estimator)*
We can theoretically build a trimmed estimator in the above framework. Let $\eta \in \mathbb{R}$ *and* $\psi(x) = \mathbb{1}_{x\geq\eta}x + \eta\mathbb{1}_{x<\eta}x$. *As* ψ *is only derivable almost everywhere, it is cautious to suppose that* f *is absolutely continuous. Then, the estimator is given by the estimating equation :*

$$\hat{\Sigma}_0 = \arg\min_{\Sigma} -\frac{1}{N}\sum_{i=1}^{N} s_\Sigma(z_i)\mathbb{1}_{f_\Sigma(z_i)>\eta} + \int s_\Sigma(z)\mathbb{1}_{f_\Sigma(z)>\eta}dz$$

In practice, the gradient descent which we will present in the next section is unable to find the global minimum. We prefer to resolve the problem for $\psi(x) = \mathbb{1}_{x\geq\eta}x + \eta\exp(x-\eta)\mathbb{1}_{x<\eta}x$

The M-estimator is Fisher-consistent :

Proposition 3.28. *Let us consider the non-normalized estimator defined by equation 3.48*

$$\hat{\Sigma}_0(P_n) = \arg\min_{\Sigma} -\int \psi(\log f_\Sigma(z))dP_n(z) + \int \Psi^*(\log f_\Sigma(z))dz \qquad (3.49)$$

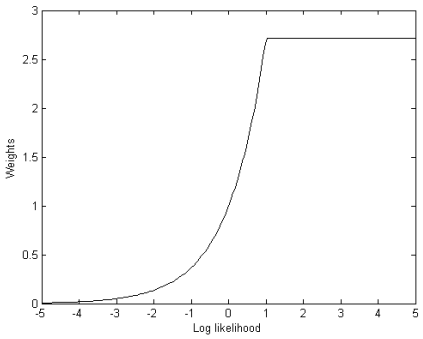


Figure 3.14 – Weighting function of the pseudo trimmed divergence

where P_n is the empirical distribution function for a sample $x_1, ..., x_n$ drawn from the distribution P_0 of density $f_{\Sigma_0}(z)$. Then

$$\hat{\Sigma}_0(P_0) = \Sigma_0$$

Proof. The result comes from a simple rewriting of the estimator as a minimum of divergence :

$$\begin{aligned}
\hat{\Sigma}_0(P_0) &= \arg\min_{\Sigma} -\int \psi(\log f_\Sigma(z))dP_0(z) + \int \Psi^*(\log f_\Sigma(z))dz \\
&= \arg\min_{\Sigma} \int \psi(\log f_{\Sigma_0}(z))dP_0(z) - \int \psi(\log f_\Sigma(z))dP_0(z) \\
&\quad + \int \Psi^*(\log f_{\Sigma_0}(z))dz - \int \Psi^*(\log f_\Sigma(z))dz \\
&= \arg\min_{\Sigma} d_\psi(f_{\Sigma_0}, f_\Sigma) = \Sigma_0
\end{aligned}$$

□

Normalized M-estimator

Fujisawa and Kano [53] considered a different divergence, close to the Rényi divergence, which is invariant by multiplication by a scalar.

Definition 3.21. *We consider a γ-divergence parametrized by $\gamma > 0$ and defined by*

$$d_\gamma(f, g) = \log\left[\frac{\left(\int g(x)^{1+\gamma}dx\right)^{1/\gamma(1+\gamma)}\left(\int f(x)^{1+\gamma}dx\right)^{1/(1+\gamma)}}{\left(\int g(x)f(x)^\gamma dx\right)^{1/\gamma}}\right] \tag{3.50}$$

Proposition 3.29. *d_γ is a divergence in the sense that for two densities functions f and g*

- *$d_\gamma(f, g) \geq 0$*
- *$d_\gamma(f, g) = 0$ if and only if $f = g$*

Furthermore, for any $\lambda_1, \lambda_2 > 0$, $d_\gamma(\lambda_1 f, \lambda_2 g) = d_\gamma(f, g)$

Proof. The two first properties are a result of the Holder inequality with equality if f and g are proportional that happens if and only if $f = g$ for two densities. The third property is straightforward from the definition of the divergence. □

This divergence leads to the following estimator

$$\hat{\Sigma}_0 = \arg\min_{\Sigma} -\frac{1}{\gamma}\log\left[\frac{1}{N}\sum_{i=1}^{N} f_\Sigma(x_i)^\gamma\right] + \frac{1}{\gamma+1}\log\left[\int f_\Sigma(x)^{1+\gamma}dx\right] \tag{3.51}$$

This leads to consider the following estimating equation

$$\frac{\sum_{i=1}^N f_{\hat{\Sigma}_0}(z_i)^\gamma s(z_i,\hat{\Sigma}_0)}{\sum_{i=1}^N f_{\hat{\Sigma}_0}(z_i)^\gamma} = \frac{\mathbb{E}\left[f_{\hat{\Sigma}_0}(z)^\gamma s(z,\hat{\Sigma}_0)\right]}{\mathbb{E}\left[f_{\hat{\Sigma}_0}(z)^\gamma\right]}$$

If we define $w_i(z_1,...,z_N,\Sigma) = \frac{f_{\hat{\Sigma}_0}(z_i)^\gamma}{\sum_{i=1}^N f_{\hat{\Sigma}_0}(z_i)^\gamma}$ and $w(z,\Sigma) = \frac{f_{\hat{\Sigma}_0}(z_i)^\gamma}{\mathbb{E}\left[f_{\hat{\Sigma}_0}(z)^\gamma\right]}$, this can be rewritten

$$\sum_{i=1}^{N} w_i(z_1,...,z_N,\hat{\Sigma}_0)s(z_i,\hat{\Sigma}_0) = \mathbb{E}\left[w(z,\hat{\Sigma}_0)s(z,\hat{\Sigma}_0)\right]$$

with the property $\sum_{i=1}^N w_i(z_1,...,z_N,\Sigma) = 1$ and $\mathbb{E}[w(z,\Sigma)] = 1$ which justifies its name of normalized estimator.

As for the non-normalized estimator, we have a Fisher-consistency result

Proposition 3.30. *Let us consider the normalized estimator defined by equation 3.48*

$$\hat{\Sigma}_0(P_n) = \arg\min_{\Sigma} -\frac{1}{\gamma}\log\left[\int f_\Sigma(x)^\gamma dP_n(x)\right] + \frac{1}{\gamma+1}\log\left[\int f_\Sigma(x)^{1+\gamma}dx\right]$$

where P_n is the empirical distribution function for a sample $x_1,...,x_n$ drawn from the distribution P_0 of density $f_{\Sigma_0}(z)$. Then

$$\hat{\Sigma}_0(P_0) = \Sigma_0$$

Proof. The result comes again from a rewriting of the estimator as a minimum of divergence :

$$\begin{aligned}\hat{\Sigma}_0(P_0) &= \arg\min_{\Sigma} -\frac{1}{\gamma}\log\left[\int f_\Sigma(x)^\gamma dP_n(x)\right] + \frac{1}{\gamma+1}\log\left[\int f_\Sigma(x)^{1+\gamma}dx\right] \\ &= \arg\min_{\Sigma} d_\gamma(f_{\Sigma_0}, f_\Sigma) = \Sigma_0\end{aligned}$$

□

Remark 3.46. *We defined the normalized and the non-normalized estimators through the minimization formulation because the roots of the estimating equations are not necessarily unique.*

The parameter γ tuned the robustness of the estimator. The estimation will be more efficient if γ is close to 0. The limit case $\gamma \to 0$ leads to the minimization of the Kullback-Leibler divergence and the maximum likelihood estimator which is the most efficient estimator.

Pythagorean relation

We consider the gross error model $g = (1-\epsilon)f_0 + \epsilon f_{out}$ where f_0 and f_{out} respectively represent the true and the outliers density.
The interest of the normalized estimation in term of robustness even if the parameter ϵ is not small can be based on the underlying Pythagorean relation expressed by Fujisawa [52]. Let us begin with the following Lemma

Lemma 3.15. *Suppose that for a density function h verifies*

$$I(h) = \left(\int f_{out}(x) h(x)_0^{\gamma} dx \right)^{1/\gamma_0} \text{ is sufficiently small for an appropriately large } \gamma_0$$

Then,

$$d_\gamma(g,h) = d_\gamma(f,h) - \frac{1}{\gamma}\log(1-\epsilon) + \epsilon O(I(h)^\gamma)$$

Proof. We produce the proof proposed by [52].

$$\begin{aligned}
d_\gamma(g,h) &= -\frac{1}{\gamma}\log\left[\int g(x)h(x)^\gamma dx\right] + \frac{1}{1+\gamma}\log\left[\int h(x)^{1+\gamma}dx\right] + A(h) \\
&= -\frac{1}{\gamma}\log\left[\int ((1-\epsilon)f_0(x) + \epsilon f_{out}(x))h(x)^\gamma dx\right] + \frac{1}{1+\gamma}\log\left[\int h(x)^{1+\gamma}dx\right] + A(h) \\
&= -\frac{1}{\gamma}\int (1-\epsilon)f_0(x)h(x)^\gamma dx + \frac{1}{1+\gamma}\int h(x)^{1+\gamma}dx + \epsilon O(I(h)^\gamma) + A(h) \\
&= d_\gamma((1-\epsilon)f,h) + \epsilon O(I(h)^\gamma) \\
&= d_\gamma(f,h) - \frac{1}{\gamma}\log(1-\epsilon) + \epsilon O(I(h)^\gamma)
\end{aligned}$$

The third and last equality comes from the Lyapunov inequality stated for $\gamma \leq \gamma_0$ (which is sufficiently large)

$$\int f_{out}(x) f_0(x)^\gamma dx \leq \left(\int f_{out}(x) f_0(x)^{\gamma_0} dx \right)^{\gamma/\gamma_0} = I(h)^\gamma$$

□

We are now ready to state the Pythagorean relation

Proposition 3.31. *Suppose that the density h verifies*

$$I(h) = \left(\int f_{out}(x) h(x)^{\gamma_0} dx \right)^{1/\gamma_0} \text{ is sufficiently small for an appropriately large } \gamma_0 \tag{3.52}$$

Then

$$d_\gamma(g,h) - d_\gamma(g,f) - d_\gamma(f,h) = \epsilon O(\max(I(h), I(f))^\gamma)$$

Proof. We deduce from the previous Lemma 3.15

$$\begin{aligned}
& d_\gamma(g,h) - d_\gamma(g,f) - d_\gamma(f,h) = \epsilon O(\max(I(h), I(f))^\gamma) \\
= \;& d_\gamma(f,h) - \frac{1}{\gamma}\log(1-\epsilon) + \epsilon O(I(h)^\gamma) + \frac{1}{\gamma}\log(1-\epsilon) + \epsilon O(I(f)^\gamma) - d_\gamma(f,h) \\
= \;& \epsilon O(\max(I(h), I(f))^\gamma)
\end{aligned}$$

□

This relation implies a projection structure. If h represents any distribution of the parametrized model (f_Σ), $f_0 = f_{\Sigma_0}$ will be the "orthogonal' projection of g onto the space described by h.

$$\arg\min_\Sigma d_\gamma(g, f_\Sigma) \approx \arg\min_\Sigma d_\gamma(f_0, f_\Sigma) + d_\gamma(g, f_0) = \Sigma_0$$

For an outlier distribution sufficiently far from the model in the sense of equation 3.52, the normalized estimator will then asymptotically focus on true parameter Σ_0 even under the gross error model.

3.6.2 Application to Angular Complex Gaussian distribution

Non-Normalized M-estimator

If we apply the non-normalized approach to the elliptical distribution, we find the estimating equation proposed by Maronna [78] as robust M-estimators for the inference of the scatter matrix.
Indeed, we can write the minimization problem :

$$\hat{\Sigma} = \arg\inf_{\Sigma} -\frac{1}{N}\sum_{i=1}^{N} \psi \circ \log\left(\frac{\Gamma(d)}{2\pi^d|\Sigma|}(x^+\Sigma^{-1}x)^{-d}\right) - \int \Psi^* \circ \log\left(\frac{\Gamma(d)}{2\pi^d|\Sigma|}(x^+\Sigma^{-1}x)^{-d}\right) dx \tag{3.53}$$

which implies the following estimating equation :

$$\begin{aligned} & -\frac{1}{N}\sum_{i=1}^{N} \psi' \circ \log\left(\frac{\Gamma(d)}{2\pi^d|\Sigma|}(z_i^+\Sigma^{-1}z_i)^{-d}\right)\left(d\frac{z_i z_i^+}{z_i^+\Sigma^{-1}z_i} - \Sigma\right) \\ = \; & \mathbb{E}\left[\psi' \circ \log\left(\frac{\Gamma(d)}{2\pi^d|\Sigma|}(z^+\Sigma^{-1}z)^{-d}\right)\left(d\frac{zz^+}{z^+\Sigma^{-1}z} - \Sigma\right)\right] \end{aligned}$$

Proposition 3.32. *Let ψ be a convex increasing and differentiable function and $x_1, ..., x_N$ a sample coming from an ACG distribution of scatter matrix Σ_0. Then the estimator 3.48 defined by*

$$\hat{\Sigma}_0 = \arg\min_{\Sigma} -\frac{1}{N}\sum_{i=1}^{N} \psi(\log f_\Sigma(z_i)) + \int \Psi^*(\log f_\Sigma(z))dz \tag{3.54}$$

is consistent i.e.

$$\hat{\Sigma}_0 \xrightarrow{p} \Sigma_0$$

Proof. We will note in the following :

$$D_n(\Sigma) = -\frac{1}{N}\sum_{i=1}^{N} \psi(\log f_\Sigma(z_i)) + \int_{S_d} \Psi^*(\log f_\Sigma(z))dz$$

and

$$D(\Sigma) = -\int_{S_d} \psi(\log f_\Sigma(x)) f_{\Sigma_0}(x)dx + \int_{S_d} \Psi^*(\log f_\Sigma(z))dz$$

The crucial part for the consistency of the estimator is the uniform convergence of the minimizer functional that is not true for all $\Sigma \in \mathcal{S}_d^{++}(\mathbb{C})$ with fixed trace. Then, we will first prove the following lemma

Lemma 3.16. *There exists a constant $m > 0$ such that*

$$Tr(\hat{\Sigma}_0^{-1}) \leq \frac{1}{m}$$

almost surely

Proof. Let us suppose the contrary. Then, there exists a subdivision φ and a collection of events Ω of non-zero measure such that

$$Tr(\hat{\Sigma}_{0,\varphi(n)}^{-1}(\omega)) \to \infty \text{ for all } \omega \in \Omega$$

We will omit the reference to n and ω in the following for the sake of clarity.
Yet for $x \in S_d$, it holds for any Σ

$$\log\left(\frac{\langle x, e_{max}\rangle Tr(\Sigma^{-1})}{d s_d (\det \Sigma^{-1})^{1/d}}\right) \leq \log f_\Sigma(x) \leq \log\left(\frac{x^+ x Tr(\Sigma^{-1})}{s_d (\det \Sigma^{-1})^{1/d}}\right)$$

where e_{max} is the eigenvector associated to the maximum eigenvalue of Σ^{-1}.
As ψ ans Ψ^* are increasing, we deduce from this inequality

$$D_n(\Sigma) \geq \int_{S_d} \Psi^*\left(\log\left(\frac{\langle x, e_{max}\rangle Tr(\Sigma^{-1})}{d s_d (\det \Sigma^{-1})^{1/d}}\right)\right) dx - \psi\left(\log\left(\frac{Tr(\Sigma^{-1})}{s_d (\det \Sigma^{-1})^{1/d}}\right)\right)$$

We can now state a property of the function Ψ^*

Lemma 3.17. *For $x > 0$,*

$$\Psi^*(x) \geq (\psi(x) - \psi(0))\frac{e^x}{x} \tag{3.55}$$

Proof. As ψ is convex, let $x > 0$ and $0 \leq t \leq x$. Then

$$\begin{aligned}
\Psi^*(x) &= \int_0^x \psi'(t) e^t dt \\
&= \psi(x)e^x - \psi(0) - \int_0^x \psi(t) e^t dt \\
&\geq \psi(x)e^x - \psi(0) - \int_0^x \left[(\psi(x) - \psi(0))\frac{t}{x} + \psi(0)\right] e^t dt \text{ by convexity of } \psi \\
&\geq \psi(x)e^x - \psi(0) - (\psi(x) - \psi(0))\frac{xe^x - e^x}{x} - \psi(0)(e^x - 1) \\
&\geq (\psi(x) - \psi(0))\frac{e^x}{x}
\end{aligned}$$

□

Furthermore, if $Tr(\Sigma) = d$ and if we note $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_d$ the eigenvalues of Σ^{-1}, we have

$$d = Tr(\Sigma) = \sum_{i=1}^d \frac{1}{\lambda_i} \geq \frac{d}{\lambda_1}$$

i.e. $\lambda_1 \geq 1$ thus

$$(\det \Sigma^{-1})^{1/d} \leq \lambda_d^{(d-1)/d} \leq \left(\frac{Tr(\Sigma^{-1})}{d}\right)^{(d-1)/d}$$

and $Tr(\Sigma^{-1})/(\det \Sigma^{-1})^{1/d} \geq d^{(d-1)/d} Tr(\Sigma^{-1})^{1/d}$ i.e. as $Tr(\hat{\Sigma}_0^{-1}) \to \infty$

$$\frac{Tr(\hat{\Sigma}_0^{-1})}{(\det \hat{\Sigma}_0^{-1})^{1/d}} \to \infty.$$

Then for n large enough, there exists a constant C such that

$$\int_{S_d} \Psi^*\left(\log\left(\frac{\langle x, e_{max}\rangle Tr(\Sigma^{-1})}{d s_d (\det \Sigma^{-1})^{1/d}}\right)\right) dx \geq C \int_{x \in S_d, \langle x, e_{max}\rangle \geq 1/2} \Psi^*\left(\log\left(\frac{\langle x, e_{max}\rangle Tr(\Sigma^{-1})}{d s_d (\det \Sigma^{-1})^{1/d}}\right)\right) dx$$

Hence,

$$D_{\varphi(n)}(\hat{\Sigma}_0) \geq C\Psi^*\left(\log\left(\frac{Tr(\hat{\Sigma}_0^{-1})}{2 d s_d (\det \hat{\Sigma}_0^{-1})^{1/d}}\right)\right) - \psi\left(\log\left(\frac{Tr(\hat{\Sigma}_0^{-1})}{s_d (\det \hat{\Sigma}_0^{-1})^{1/d}}\right)\right) \tag{3.56}$$

By combining Equations 3.55 and 3.56, we get $D_{\varphi(n)}(\hat{\Sigma}_0) \to \infty$ which proves the contradiction. □

We can now refine the space of parameters

$$M = \left\{ \Sigma \in \mathcal{S}_d^{++}(\mathbb{C}) \text{ such that } Tr(\Sigma) = d \text{ and } Tr(\Sigma^{-1}) \leq \frac{1}{m} \right\}$$

For this space, we prove that

Lemma 3.18.

$$\sup_{\Sigma \in M} |D_N(\Sigma) - D(\Sigma)| \overset{a.s.}{\to} 0$$

Proof. Let note $\xi(x, \Sigma) = \psi(\log f_\Sigma(x))$ for $x \in S_d$. Then

$$|D_n(\Sigma) - D(\Sigma)| = \left| \frac{1}{N} \sum_{i=1}^{N} \xi(x_i, \Sigma) - \int_{S_d} \xi(x, \Sigma) f_{\Sigma_0}(x) dx \right|.$$

The result is an application of the uniform law of large numbers since

- M is compact because $Tr(\Sigma^2) \leq d^2$ for $\Sigma \in M$
- $\Sigma \mapsto \xi(x, \Sigma)$ is continuous on M for all $x \in S_d$
- $\xi(x, \Sigma) \leq \psi\left(\log\left(\frac{d^{d+1} Tr(\Sigma^{-1})}{s_d}\right)\right) \leq C'$ where C' is a constant.

□

Finally, it remains to prove a coercive condition of the optimum of $D(\Sigma)$. For that purpose, it suffices to remark that M is compact and

$$D(\Sigma) = d_\psi(\Sigma, \Sigma_0) + D(\Sigma_0) \tag{3.57}$$

where d_ψ is a ψ-divergence. Then, for any $\epsilon > 0$,

$$\inf_{\Sigma \in M \text{ s.t. } Tr(\Sigma - \Sigma_0)^2 > \epsilon} D(\Sigma) = \min_{\Sigma \in M \text{ s.t. } Tr(\Sigma - \Sigma_0)^2 > \epsilon} D(\Sigma) > D(\Sigma_0)$$

We can apply Theorem 5.7 on M-estimators proposed by Van der Vaart [106]. □

Normalized M-estimator

The same process can be applied for normalized M-estimator :

$$\hat{\Sigma} = \arg\inf_{\Sigma} -\frac{1}{\gamma} \log \left[\frac{1}{N} \sum_{i=1}^{N} (z_i^+ \Sigma^{-1} z_i)^{-d\gamma} \right] + \frac{1}{\gamma + 1} \log \left[\int (x^+ \Sigma^{-1} x)^{-d(1+\gamma)} dx \right] \tag{3.58}$$

which implies the following estimating equation :

$$-\frac{1}{N} \sum_{i=1}^{N} w_i(z_1, ..., z_N, \Sigma) \frac{z_i z_i^+}{z_i^+ \Sigma^{-1} z_i} = \mathbb{E}\left[w(z) \frac{z z^+}{z^+ \Sigma^{-1} z} \right]$$

with

$$w_i(z_1, ..., z_N, \Sigma) = \frac{(z_i^+ \Sigma^{-1} z_i)^{-d\gamma}}{\sum_{i=1}^{N} (z_i^+ \Sigma^{-1} z_i)^{-d\gamma}}$$

$$w(z, \Sigma) = \frac{(z^+ \Sigma^{-1} z)^{-d\gamma}}{\mathbb{E}\left[(z^+ \Sigma^{-1} z)^{-d\gamma}\right]}$$

Proposition 3.33. *Let $\gamma > 0$, $x_1, ..., x_N$ an iid sample coming from an ACG distribution of scatter matrix Σ_0. Then the estimator defined by*

$$\hat{\Sigma}_0 = \arg\min_{\Sigma} -\frac{1}{\gamma}\log\left[\frac{1}{N}\sum_{i=1}^{N} f_\Sigma(x_i)^\gamma\right] + \frac{1}{\gamma+1}\log\left[\int f_\Sigma(x)^{1+\gamma}dx\right] \tag{3.59}$$

is consistent i.e.

$$\hat{\Sigma}_0 \xrightarrow{p} \Sigma_0$$

Proof. We follow the same sketch of proof as in the previous consistency Proposition 3.32. We first note that there exists $m > 0$ such that

$$Tr(\hat{\Sigma}_0^{-1}) \leq \frac{1}{m}$$

by the same lower bound argument. We can then define the same parameter space :

$$M = \left\{\Sigma \in \mathcal{S}_d^{++}(\mathbb{C}) \text{ such that } Tr(\Sigma) = d \text{ and } Tr(\Sigma^{-1}) \leq \frac{1}{m}\right\}$$

We note

$$D_N(\Sigma) = -\frac{1}{\gamma}\log\left[\frac{1}{N}\sum_{i=1}^{N} f_\Sigma(x_i)^\gamma\right] + \frac{1}{\gamma+1}\log\left[\int f_\Sigma(x)^{1+\gamma}dx\right]$$

and

$$D(\Sigma) = -\frac{1}{\gamma}\log\left[\int_{S_d} f_\Sigma(x)^\gamma f_{\Sigma_0}(x)dx\right] + \frac{1}{\gamma+1}\log\left[\int f_\Sigma(x)^{1+\gamma}dx\right]$$

Then

Lemma 3.19.

$$\sup_{\Sigma\in M} |D_N(\Sigma) - D(\Sigma)| \xrightarrow{a.s.} 0$$

Proof. For $x \in S_d$ and $\Sigma \in M$, we have

$$\frac{x^+ x\lambda_{min}}{s_d(\det\Sigma^{-1})^{1/d}} \leq f_\Sigma(x) \leq \frac{x^+ x Tr(\Sigma^{-1})}{s_d(\det\Sigma^{-1})^{1/d}}$$

where λ_{min} is the minimum eigenvalue of Σ^{-1} i.e.

$$0 < \frac{dm}{s_d} \leq \frac{d^2}{s_d Tr(\Sigma)Tr(\Sigma^{-1})} \leq f_\Sigma(x) \leq \frac{Tr(\Sigma^{-1})Tr(\Sigma)}{s_d} \leq \frac{d}{s_d m}$$

i.e. $f_\Sigma(x)$ is almost surely bounded. Then almost surely

$$\sup_{\Sigma\in M} |D_N(\Sigma) - D(\Sigma)| \leq \left(\sup_{\frac{dm}{s_d}<t<\frac{d}{s_d m}} \frac{1}{t}\right) \sup_{\Sigma\in M}\left|\frac{1}{N}\sum_{i=1}^{N} f_\Sigma(x_i)^\gamma - \int_{S_d} f_\Sigma(x)^\gamma f_{\Sigma_0}(x)dx\right| \to \infty$$

by applying the uniform law of large numbers. □

We end the proof with Theorem 5.7 of van der Vaart [106]. □

Influence functions

The influence function is a classical tool used to evaluate the robustness of an estimator. It measures the impact of an outlier in the estimation process. An estimator is said to be robust if the influence function is bounded whenever the value of the outlier.

Non-normalized M-estimator

Let θ be the vectorization of Σ ($\theta := \text{vec}(\Sigma)$). If we denote for a cumulative distribution function P by $\hat{\theta}_\psi(P)$ the minimizer of the cross entropy :

$$\hat{\theta}_\psi(P) = \arg\min_\theta - \int \psi(\log f_\theta(z))dP(z) + \int \Psi^*(\log f_\theta(z))dz \tag{3.60}$$

the influence function is defined for the "outlier" $y \in \mathbb{C}^d$ by

$$IF(y,\theta) = \left.\frac{\partial \hat{\theta}_\psi(P_\epsilon)}{\partial \epsilon}\right|_{\epsilon=0}$$

We recall that $P_\epsilon = (1-\epsilon)P_\theta + \epsilon\delta_y$ corresponds to the distribution function of the gross error model. δ_y designates the Dirac distribution at $y \in S_d$.
If we write $\hat{\theta}_\epsilon = \hat{\theta}_\psi(P_\epsilon)$, the estimating equation becomes :

$$\int \psi'(\log f_{\hat{\theta}_\epsilon}(z))s(z,\hat{\theta}_\epsilon)dP_\epsilon(z) = \int \psi'(\log f_{\hat{\theta}_\epsilon}(z))s(z,\hat{\theta}_\epsilon)dP_{\hat{\theta}_\epsilon}(z).$$

As the definition of the influence function implies a differentiating, the multiplication constant of the estimator $\hat{\theta}_\epsilon = \text{vec}(\hat{\Sigma}_\epsilon)$ is crucial and leads to different expressions. In the following, we will assume that

$$Tr(\hat{\Sigma}_\epsilon) = d$$

Then, we have the following proposition

Proposition 3.34. *The influence function of the non-normalized estimator $\hat{\theta} = vec(\hat{\Sigma})$ under the constraint $Tr(\hat{\Sigma}) = d$ is given by*

$$IF(y,\theta) = (\Pi(J(\theta)+I_{d^2})\Pi - I_{d^2})^{-1}\,\Pi\left(\psi'(\log f(y,\theta))s(y,\theta) - \int \psi'(\log f_\theta(z))s(z,\theta)dP_\theta(z)\right) \tag{3.61}$$

with

$$\begin{cases} s(z,\theta) = vec\left(\Sigma^{-1}\frac{dzz+}{z+\Sigma^{-1}z}\Sigma^{-1} - \Sigma^{-1}\right) \\ J(\theta) = \int \psi'(\log f_\theta(z))s(z,\theta)s(z,\theta)^+dP_\theta(z) \\ \Pi = I_{d^2} - \frac{1}{d}vec(I_d)vec(I_d)^+ \end{cases}$$

Remark 3.47. *$s(z,\theta)$ is the score function with respect to θ. It is the vectorization of the score function with respect to Σ :*

$$s(z,\Sigma) = \Sigma^{-1}\frac{dzz+}{z+\Sigma^{-1}z}\Sigma^{-1} - \Sigma^{-1}$$

Proof. If we differentiate the estimating equation with respect to ϵ and we take $\epsilon \to 0$, we obtain

$$J(\theta)IF(y,\theta) = \left(\psi'(\log f(y,\theta))s(y,\theta) - \int \psi'(\log f_\theta(z))s(z,\theta)dP_\theta(z)\right)$$

It is clear that $J(\theta)$ is badly conditioned because of the constraint $Tr(\hat{\Sigma}_\epsilon) = m$ that can be written for $\hat{\theta}_\epsilon$:

$$\hat{\theta}_\epsilon^+ i_d = 1$$

with $i_d = \text{vec}(I_d)$. We differentiate this equality with respect to ϵ which gives us

$$IF(y,\theta)^+ i_d = 0$$

$\Pi = I_{d^2} - \frac{1}{d}\text{vec}(I_d)\text{vec}(I_d)^+$ is the projector on the orthogonal space of i_d. Then the solution given by

$$\left(\Pi(J(\theta) + I_{d^2})\Pi - I_{d^2}\right) IF(y,\theta) = \Pi\left(\psi'(\log f(y,\theta))s(y,\theta) - \int \psi'(\log f_\theta(z))s(z,\theta)dP_\theta(z)\right)$$

verifies

$$\begin{aligned} i_d^+ IF(y,\theta) &= -i_d^+ \left(\Pi(J(\theta) + I_{d^2})\Pi - I_{d^2}\right) IF(y,\theta) \\ &= -i_d^+ \Pi \left(\psi'(\log f(y,\theta))s(y,\theta) - \int \psi'(\log f_\theta(z))s(z,\theta)dP_\theta(z)\right) = 0 \end{aligned}$$

□

Figures 3.15 and 3.16 present the Frobenius norm of influence functions for $\Sigma = \begin{pmatrix} 1 & -\mu \\ -\overline{\mu} & 1 \end{pmatrix}$ and different ψ. We consider outliers as pure frequencies

$$y = \frac{[1; e^{2i\pi\nu}; e^{2i\pi(2\nu)}; \ldots; e^{2i\pi(d-1)\nu}]}{\sqrt{d}}$$

These outliers represent ideal targets in radar applications. This allows us to plot the norm of influence function for different frequencies, covariance and function ψ.

Normalized M-estimator

We note $w(z,\theta) = (z^+\Sigma^{-1}z)^{-d\gamma}$ and $V(z,\theta) = d(z^+\Sigma^{-1}z)^{-d\gamma}\text{vec}(\frac{\Sigma^{-1}zz^+\Sigma^{-1}}{z^+\Sigma^{-1}z})$.
By performing the same operations on the normalized estimating equation, we find :

Proposition 3.35. *The influence function of the normalized estimator $\hat{\theta} = vec(\hat{\Sigma})$ under the constraint $Tr(\hat{\Sigma}) = d$ is given by*

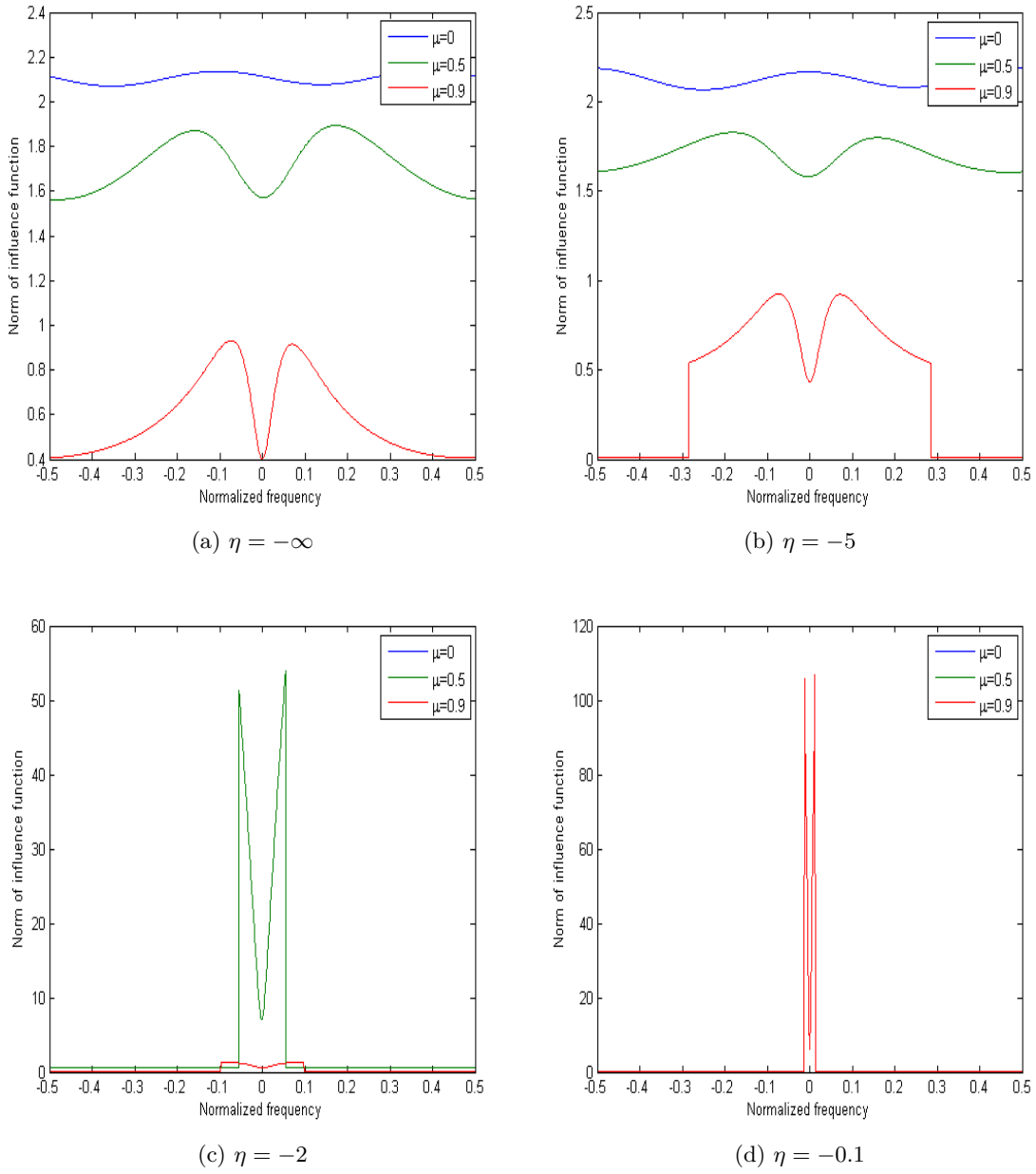
$$IF(y,\theta) = \left(\Pi(J(\theta) + I_{d^2})\Pi - I_{d^2}\right)^{-1}\left(w(y,\theta)\int V(z,\theta)dP_\theta(z) - V(y,\theta)\int w(z,\theta)dP_\theta(z)\right) \tag{3.62}$$

with

$$J(\theta) = \left(\int \frac{1}{w(z,\theta)}V(z,\theta)V(z,\theta)^+ dP_\theta(z)\right)\int w(z,\theta)dP_\theta(z) - \left(\int V(z,\theta)dP_\theta(z)\right)\left(\int V(z,\theta)dP_\theta(z)\right)^+$$

We recall that Π is the projector $\Pi = I_{d^2} - \frac{1}{d}vec(I_d)vec(I_d)^+$.

The influence of the normalized estimator illustrated by Figure 3.17 is similar to the corresponding non-normalized estimator even if thinner. β and γ have the same signification. These parameters control the robustness of the estimators as illustrated by these figures.
For $\mu = 0$, the influence function is always constant because in that case, the likelihood itself is constant. A single outlier has then no incidence on the estimation.



(a) $\eta = -\infty$ (b) $\eta = -5$

(c) $\eta = -2$ (d) $\eta = -0.1$

FIGURE 3.15 – Influence functions norm of non-normalized estimator for $\psi'(x) = \mathbb{1}_{x \geq \eta}$

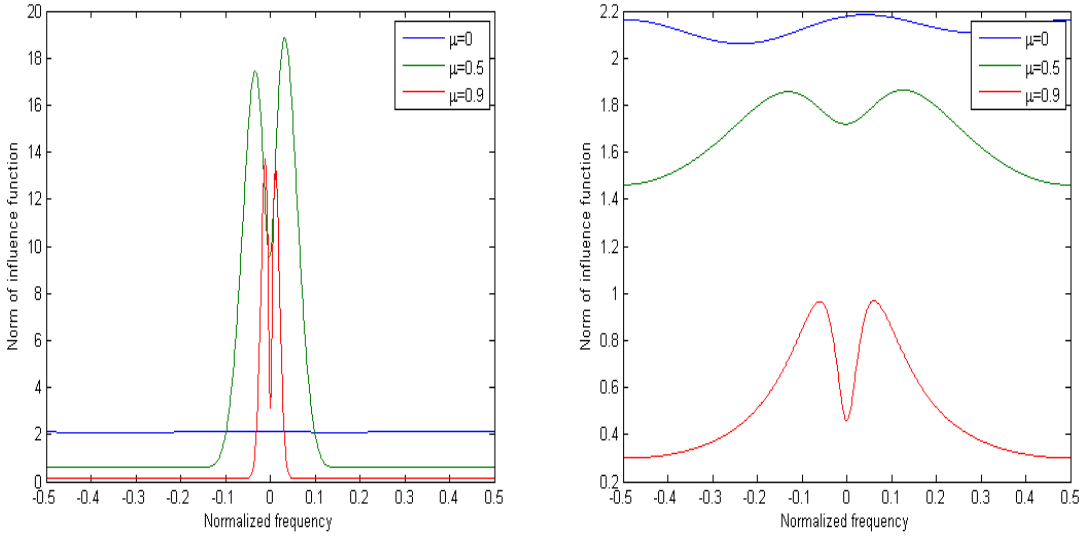
Remark 3.48. *We have implicitely given the influence function of the Tyler's estimator* $\hat{\theta} = vec(\hat{\Sigma})$ *for the constraint* $Tr(\hat{\Sigma}) = d$ *:*

$$IF(y, \theta) = \left(\Pi(J(\theta) + I_{d^2})\Pi - I_{d^2}\right)^{-1} \Pi s(y, \theta)$$

with

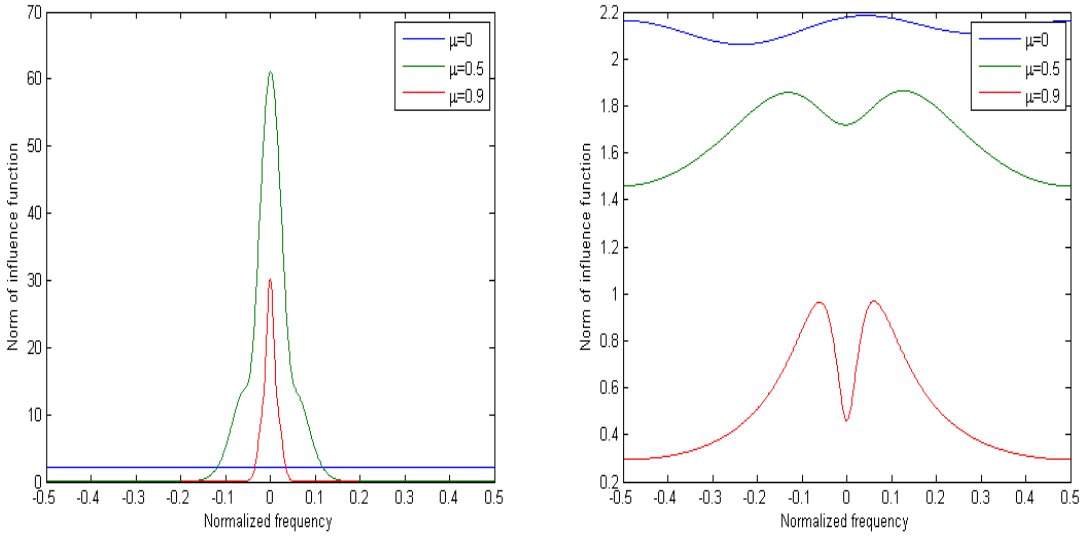
$$\begin{cases} s(z,\theta) = vec\left(\Sigma^{-1}\frac{dzz+}{z+\Sigma^{-1}z}\Sigma^{-1} - \Sigma^{-1}\right) \\ J(\theta) = \int s(z,\theta)s(z,\theta)^+ dP_\theta(z) \\ \Pi = I_{d^2} - \frac{1}{d}vec(I_d)vec(I_d)^+ \end{cases}$$

The expression is different from the one proposed by Mahot [77] for the constraint $Tr(\Sigma^{-1}\hat{\Sigma}) =$



(a) $\beta = 10$ (b) $\beta = 0.1$

Figure 3.16 – Influence functions norm of non-normalized estimator for $\psi'(x) = \exp(\beta x)$



(a) $\gamma = 10$ (b) $\gamma = 0.1$

Figure 3.17 – Influence functions norm of normalized estimator for different γ

constant

$$IF(y,\theta) = (d+1)\left(\frac{1}{d}\Sigma - \frac{yy+}{y^+\Sigma^{-1}y}\right)$$

However the constraint $Tr(\Sigma^{-1}\hat{\Sigma}) = cst$ *includes the unknown parameter* Σ.

An other measure of robustness is the robustness to the mixture of Angular Complex Gaussian distributions. We will illustrate this robustness through the stationary case in the next sections.

3.7 Autoregressive modelization and robustness with respect to heavy contamination

We introduce some methods in order to face heavy contamination in the case of an elliptical modelization. We could adapt the approach which we took in Section 3.6 to Burg methods. Indeed, we saw that we express the estimation of each reflection parameter as the minimization of a functional error depending on the error vectors b_m and f_m. As the vectors $(b_m(n-1), f_m(n))^T$ have an elliptical distribution, we could estimate their correlation parameter through the minimization of a divergence defined in Section 3.6.
We present in Section 3.7.2 an other approach that can attain the same objective i.e. robustify the estimation of the reflection parameters.

3.7.1 Burg M-estimator

The estimators of a reflection parameter could have been defined by :

$$\hat{\mu}_m = \arg\inf_{\mu\in D} -\frac{1}{N(m+1-d)}\sum_{i=1}^{N}\sum_{n=m+1}^{d} \psi\circ\log\left(f_\mu\left(\begin{array}{c} f_{i,m}(n) \\ b_{i,m}(n-1)\end{array}\right)\right) - \int \Psi^*\circ\log\left(f_\mu(x)\right)dx \tag{3.63}$$

with for $x\in\mathbb{C}^2$

$$f_\mu(x) = \frac{1}{2\pi^2(1-|\mu|^2)}\left(x^+\begin{pmatrix} 1 & \mu \\ \overline{\mu} & 1\end{pmatrix}x\right)^{-2}$$

or

$$\begin{aligned}\hat{\mu}_m = \arg\inf_{\mu\in D} \quad & -\frac{1}{\gamma}\log\left[\frac{1}{N(m+1-d)}\sum_{i=1}^{N}\sum_{n=m+1}^{d}\left(\begin{pmatrix} f_{i,m}(n) \\ b_{i,m}(n-1)\end{pmatrix}^+\begin{pmatrix} 1 & \mu \\ \overline{\mu} & 1\end{pmatrix}\begin{pmatrix} f_{i,m}(n) \\ b_{i,m}(n-1)\end{pmatrix}\right)^\gamma\right] \\ & + \frac{1}{\gamma+1}\log\left[\int\left(x^+\begin{pmatrix} 1 & \mu \\ \overline{\mu} & 1\end{pmatrix}x\right)^{1+\gamma}dx\right].\end{aligned}$$

Unfortunately, this method is not well adapted because the Burg technique implies an iterative procedure. Indeed, the ACG distribution has difficulties to estimate the scatter matrix Σ for a contaminated sample if the true matrix is $\Sigma = I_d$. This is due to the definition of an outlier for an ACG distribution illustrated by the figure 3.4b. In case of a true distribution of scatter matrix equal to I_d, every sample has the same weight.
Furthermore, as the Burg technique iteratively estimate a 2×2 correlation parameter, we often face the case where $\mu_m = 0$ i.e. the covariance of the error vectors $\begin{pmatrix} f_{i,m}(n) \\ b_{i,m}(n-1)\end{pmatrix}$ is proportional to I_2, especially in radar applications.

3.7.2 Geodesic Burg estimators

The idea of Geodesic Burg estimators is to cut the samples $x_1, ..., x_N$ into S subsamples and to perform on each subsample an estimation process such as a classical Burg estimator or Normalized Burg estimators (preferably fast but not necessarily robust). Then, we will compute a "representative" for the S estimates $\hat{\mu}_1, ..., \hat{\mu}_S$.

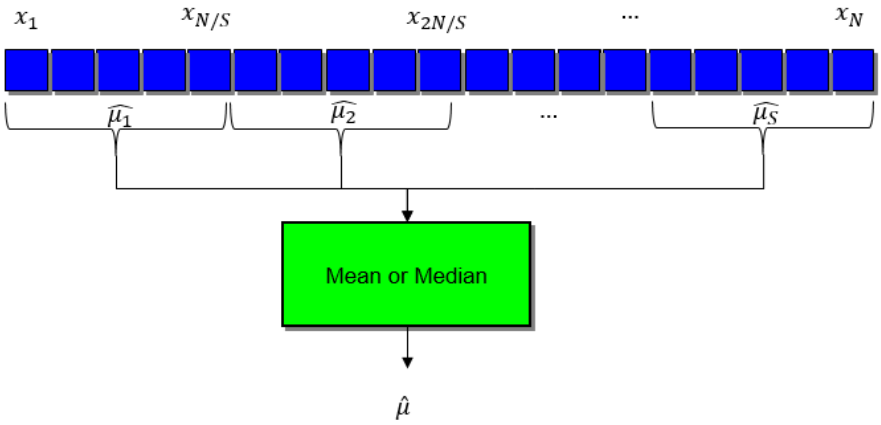


FIGURE 3.18 – Splitting of the sample $x_1, ..., x_N$ into S subsamples

As representative, we will consider the mean or median of $\hat{\mu}_1, ..., \hat{\mu}_S$ defined by the following minimization problems

$$\hat{\mu}_{mean} = \arg\min_{\mu} \sum_{i=1}^{S} d(\mu, \hat{\mu}_i)^2$$

and

$$\hat{\mu}_{median} = \arg\min_{\mu} \sum_{i=1}^{S} d(\mu, \hat{\mu}_i)$$

where d is a distance defined on the Poincaré disk $D = \{z \in \mathbb{C} \text{ s.t. } |z| < 1\}$. It is natural to consider a Riemannian distance that reflects the geometry of the model. As we saw in Section 3.3, a natural metric to consider in the Poincaré disk is the Poincaré metric. Recall that it corresponds to the geometry related to the entropy of an autoregressive vector, whose distance is given by

$$d(\mu_1, \mu_2) = \frac{1}{2} \log \frac{1 + \left| \frac{\mu_1 - \mu_2}{1 - \overline{\mu_1} \mu_2} \right|}{1 - \left| \frac{\mu_1 - \mu_2}{1 - \overline{\mu_1} \mu_2} \right|}$$

These mean and median computations for this metric were discussed by Arnaudon et al. [6] and especially the median computation in its generality in Riemannian spaces was studied by Yang [111]. They gave two algorithms slightly simpler than steepest descent algorithms proposed previously.
Let us recall that the exponential map associated with the distance d in the Poincaré disk is given by

$$\exp_{\mu}(v) = \frac{(\mu + e^{i\theta}) e^{2|v|_D} + (\mu - e^{i\theta})}{(1 + \overline{\mu} e^{i\theta}) e^{2|v|_D} + (1 - \overline{\mu} e^{i\theta})}$$

where $\theta = \arg(v)$ and $|v|_D = \frac{|v|}{1 - |\mu|^2}$.

Remark 3.49. *The limit case $S = N$ is intuitively the most robust estimator of this geodesic class. However, it corresponds to an estimation of the scatter matrix of the autoregressive vector with a single sample. It can however otherwise be explained in the framework*

Algorithm 10 Mean in the Poincaré disk

Aim : To compute the mean of $\mu_1, ..., \mu_S$ in the Poincare disk
Input : z_0 Initial state for $t = 0$, a descent step α, a tolerance η
while $|\delta_t| > \eta$
$\delta_t = \frac{1}{S} \sum_{i=1}^{S} \exp_{z_t}^{-1}(\mu_i)$
$z_{t+1} = exp_{z_t}(-\alpha\delta_t)$
end

Algorithm 11 Median in the Poincaré disk

Aim : To compute the median of $\mu_1, ..., \mu_S$ in the Poincaré disk
Input : z_0 Initial state for $t = 0$, a descent step α, a tolerance η
while $|\delta_t| > \eta$
$\delta_t = \frac{1}{S} \sum_{i=1, \mu_i \neq z_t}^{S} \frac{\exp_{z_t}^{-1}(\mu_i)}{d(z_t, \mu_i)}$
$z_{t+1} = exp_{z_t}(-\alpha\delta_t)$
end

of autoregressive processes. If the order of the underlying autoregressive M is small with respect to d, then the regularization process we presented can be interpreted as an implicit estimation of the true order of the autoregressive process [10]. The regularization is then essential in this context.

3.8 Radar detection for non-Gaussian noise

3.8.1 Test of hypotheses

We address the problem of detecting a target $p \in \mathbb{C}^d$ in a non-Gaussian noise $c \in \mathbb{C}^d$ that represents the clutter. The target vector is considered deterministic. It will be parametrized by a frequency θ :

$$p(\theta) = P_0 e^{i\alpha} \begin{pmatrix} 1 \\ e^{2i\pi\theta} \\ \vdots \\ e^{2i\pi(d-1)\theta} \end{pmatrix}$$

where P_0 represents the power and α the initial phase of the target.
Let $z \in \mathbb{C}^d$ be the signal received on a space case. The problem of detection is to decide between the two hypotheses

$$\begin{cases} H_0 : z \overset{d}{=} c \\ H_1 : z \overset{d}{=} p + c \end{cases}$$

In a non-Gaussian context, c is modelized by a centered elliptical distribution parametrized by a scatter matrix Σ and an amplitude distribution τ. We suppose that $Tr(\Sigma) = d$ and that the signal power is represented by $|\tau|^2$.
In the following, we will propose test detectors classical in the radar literature.

3.8.2 GLRT detector

In that context, the detector

$$S_\theta(z) = \frac{|p(\theta)^+\Sigma^{-1}z|^2}{(z^+\Sigma^{-1}z)(p(\theta)^+\Sigma^{-1}p(\theta))}$$

has desirable properties and has been independently derived by several authors [72] [55] [40]. We will refer to it as GLRT (Generalized Likelihood Ratio Test) but it has many names such NMF (Normalized Match Filter) or MSD (Matched Subspace Detector). We remark that the detector is independent by multiplication of p, Σ or z by a scalar. The test against a composite hypothesis H_1 where θ is unknown will be

$$S(z) = \max_{-1/2\leq\theta\leq 1/2} S_\theta(z) \tag{3.64}$$

The adaptative version of the detector for an estimator $\hat{\Sigma}$ (drawn from surrounding space cases $z_1, ..., z_N$) replacing Σ will be used in the following simulations.

3.8.3 Capon detector

The Capon detector is closer to the traditional FFT detector than the GLRT. For a discretized signal $z_1, ..., z_d$ and a filter of size $M < d$, $h = (h_1, ..., h_M)$, the filter response power is given by

$$\left|\sum_{k=1}^{M} h_k z_{n+k}\right|^2 .$$

If we consider a filter where $h_k = e^{-2i\pi k\theta}$ for $1 \leq k \leq d$ and θ a normalized frequency, we consider the empirical response to the pure frequency θ independently of the signal z. The idea of the Capon method [33] is to design an adaptive filter h for each frequency θ "as selective as possible" for the random input z.

Capon spectrum in the general case

If we denote the response of a filter h by

$$R_h(z) = \left|\sum_{k=1}^{d} h_k z_{n+k}\right|^2 ,$$

The expectation of this output is

$$\mathbb{E}[R_h(z)] = \mathbb{E}\left[\left|\sum_{k=1}^{d} h_k z_k\right|^2\right] = h^+\Sigma_M h$$

where Σ_M is the covariance of the random vector $z \in \mathbb{C}^M$. The output of the filter h for a pure frequency θ input signal is $h^+p(\theta)$ for

$$p(\theta) = \begin{pmatrix} 1 \\ e^{2i\pi\theta} \\ \vdots \\ e^{2i(M-1)\pi\theta} \end{pmatrix} .$$

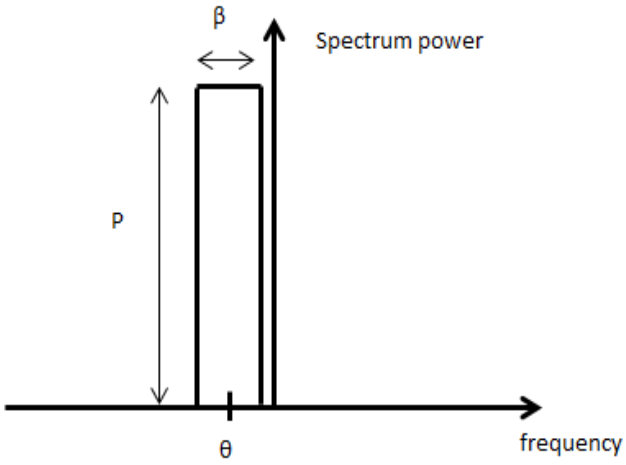


FIGURE 3.19 – Power spectral density of the filter $h_{Capon}(\theta)$ of total power $\mathbb{E}[R_h(z)]$ and bandwidth β

We want to make the filter as selective as possible for the frequency θ. Capon's idea is to maximize the expected output $\mathbb{E}[R_h]$ while the response to the frequency θ remains normalized to 1.

$$h_{Capon}(\theta) = \arg \min_{h\in\mathbb{C}^d, h^+p(\theta)=1} h^+\Sigma_M h = \frac{\Sigma_M^{-1}p(\theta)}{p(\theta)^*\Sigma_M^{-1}p(\theta)}$$

$h_{Capon}(\theta)$ corresponds to a passband filter around θ. We can then approximate the power spectral density of the input signal $S(\theta)$ by a rectangle centered on θ with a bandwidth equal to β.

Then, the value of the power spectral density at θ is approximated by

$$P = \frac{\mathbb{E}[R_h(z)]}{\beta}.$$

We will choose the simple choice for $\beta = \frac{1}{M+1}$. Thus the Capon spectrum at the frequency θ is given by

$$S_{Capon}(\theta) = \frac{M+1}{p(\theta)^+\Sigma_M^{-1}p(\theta)} \tag{3.65}$$

Capon spectrum for an autoregressive model

Let us recall that an autoregressive model of order M is parametrized by $a_1^{(M)}, ..., a_M^{(M)}$ and a power σ_M of the noise. The autoregressive vector $z_1, ..., z_d$ satisfies

$$z_n = \sum_{k=1}^{M} a_k^{(M)} z_{n-k} + \sigma_M e_n$$

with e_n a noise of variance 1 and the convention $z_n = 0$ is $n \leq 0$.

The associated power spectral density of an autoregressive model is [101]

$$S_{AR}(\theta) = \frac{\sigma_M^2}{\left|1 + \sum_{k=1}^{M} a_k^{(M)} e^{-2i\pi k\theta}\right|^2}$$

Some calculus leads to the expression of the Capon spectrum of the autoregressive of order M in function of the autoregressive spectrum of all order inferior to M

$$S_{Capon}(\theta) = \frac{M+1}{\sum_{m=0}^{M} \frac{1}{S_{AR}(\theta)}}$$

This can be equivalently written

$$S_{Capon}(\theta)^{-1} = \frac{1}{M+1} \sum_{m=0}^{M} S_{AR}(\theta)^{-1}$$

Capon detector

Let us suppose that we have the estimation of the Capon spectrum of the vector under test $\theta \mapsto S_z(\theta)$ and of the ambiance samples $\theta \mapsto S_{amb}(\theta)$. Then, the detector is

$$S = \max_{-1/2 \le \theta \le 1/2} 10\log_{10}(S_z(\theta)) - 10\log_{10}(S_{amb}(\theta)) \tag{3.66}$$

Remark 3.50. *To gain time, we can reduce the interval for the maximum defined above. In practice, for an autoregressive modelization, we take the maximum over the roots of the autoregressive polynomial*

$$A_M(z) = 1 + \sum_{k=1}^{M} a_k^{(M)} z^k$$

These roots represent the natural frequencies of the autoregressive model.

3.9 Applications for radar detection

As an illustration of the performances of the algorithms, we present a simulation of the model with a Weibull texture for the clutter. We recall the expression of the density for a Weibull distribution :

$$\text{for } x \ge 0,\ f_\tau(x) = \frac{\nu}{\sigma}\left(\frac{x}{\sigma}\right)^{\nu-1} e^{-(x/\sigma)^\nu}$$

The scale parameter σ of the texture represents the clutter power level (set to 20 dB) whereas ν is the shape parameter representing the disparity of the distribution. We will take $N = 64$ samples and a speckle built from an autoregressive vector of order 1 or 3 and of dimension $d = 8$. An $AR(1)$ is a good approximation for a radar ground clutter or a wind clutter with a single Doppler frequency whereas the $AR(3)$ models a mixture of frequencies.

3.9.1 Quality of the estimation

The measure of the error of estimation for the covariance matrix has to be normalized because the scatter matrices are defined up to a multiplicative constant. The Riemannian mean error for N estimations is the Riemannian normalized distance which is a natural distance in the space of positive definite matrices :

$$\text{MCE} = \frac{1}{N}\sum_{i=1}^{N}\left\|\log\left(\left(\frac{\hat{\Sigma}_i}{\|\hat{\Sigma}_i\|}\right)^{-1/2}\frac{\Sigma_0}{\|\Sigma_0\|}\left(\frac{\hat{\Sigma}_i}{\|\hat{\Sigma}_i\|}\right)^{-1/2}\right)\right\|_F \tag{3.67}$$

where $\|M\|_F = Tr(MM^+)$. We will consider as reference the estimators of the scatter matrix classically used in the literature

- **Empirical covariance** : given by $\hat{\Sigma}_N = \sum_{i=1}^{N} x_i x_i^+$
- **Fixed Point (FP)** : the M-estimator proposed by Tyler [104] solution of

$$\hat{\Sigma}_N = \frac{d}{N}\sum_{i=1}^{N}\frac{x_i x_i^+}{x_i^+\hat{\Sigma}_N^{-1}x_i}$$

The first comparisons will be performed with non robust (Burg type) algorithms :

- **Normalized Burg** : estimator given by equation (3.23)
- **Log Burg** : estimator given by equation (3.26) computed through a Riemannian steepest descent with a step $\alpha = 0.5$
- **Elliptical Burg** : estimator given by equation (3.29) computed through a Riemannian steepest descent with a step $\alpha = 0.5$

Secondly, we will make the comparison with the following robust algorithms

- **Geodesic Median Burg** : the covariance of the median of each estimated autoregressive model of each x_i (see for example [12] and Remark 3.49)
- **Median of Normalized Burg** : the covariance of the median of each estimated autoregressive model for $S = 8$ subsamples (equation 3.7.2)
- **ACG Non-Normalized** estimator : the scatter matrix is estimated by the elliptical non-normalized estimator (equation 3.53) for $x \mapsto \psi'(x) = e^{\beta x}$. We choose $\beta = 0.5$.
- **ACG Normalized** estimator : the scatter matrix is estimated by the elliptical normalized estimator (equation 3.58) with $\gamma = 1$.

The Burg estimators of the scatter matrix are used for a maximal order $M = d - 1$. Figure 3.20 illustrates the robustness of the different estimators with respect to the shape parameter of the Weibull texture. This robustness holds except for the empirical covariance as expected.

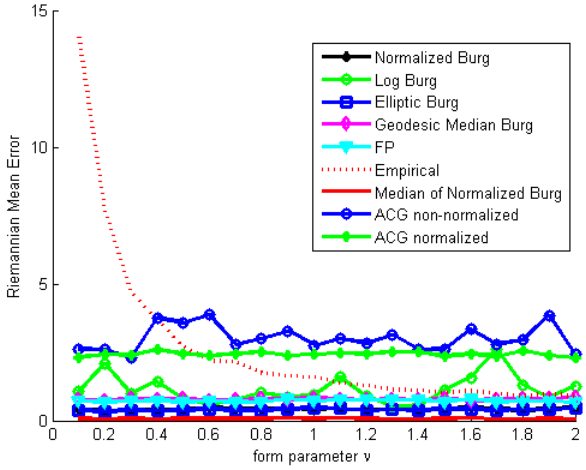


Figure 3.20 – Riemannian Mean Error for an AR(1) ($\mu_1 = 0.9$) with respect to the shape parameter of the Weibull texture

We will now present the performances of detection of the different detectors for three scenarios with the OS-CFAR (for Order Statistics- Constant False Alarm Rate) detector, classically presented in the radar literature as robust with respect to contaminating targets. Figure 3.21 presents the principle of this detector ; see [92] for details.

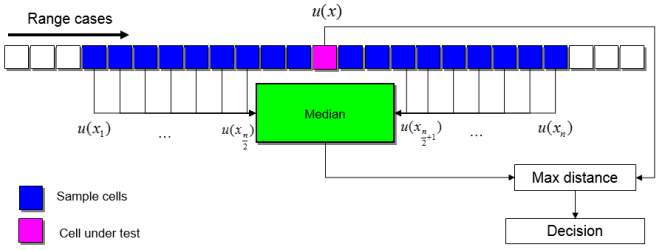


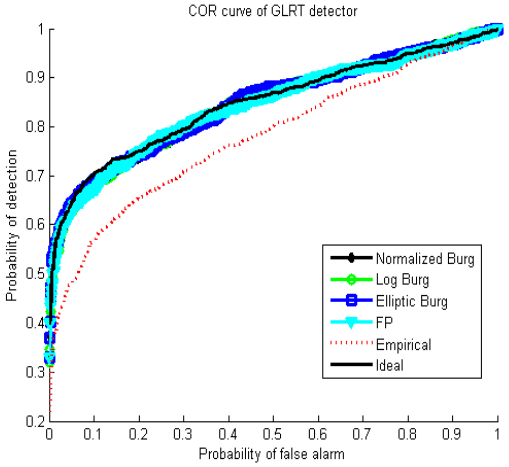
Figure 3.21 – Principle of the OS-CFAR detector : the function $u : \mathbb{C}^d \to \mathbb{R}^p$ represents p filter outputs

For more realism, we simulate a target as an autoregressive vector of order 1 with reflection parameter $\mu_1 = 0.9e^{2i\pi f}$ (where f is its normalized frequency) and P_0 that represents its power.

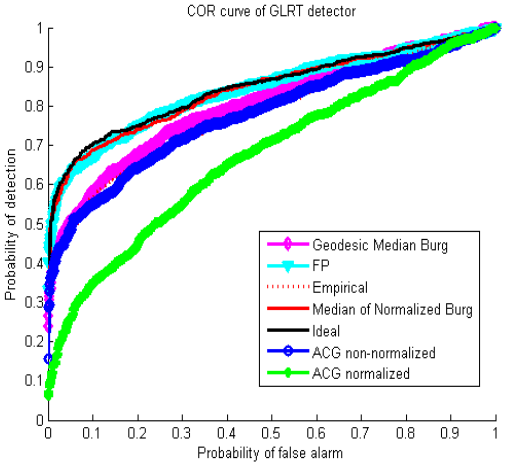
3.9.2 Scenario 0 : no outlier

First, we present the performances of the different detectors for a clutter without outliers. Figures 3.22 and 3.25 represent the COR curves for a target of power respectively equal to $20dB$ and $40dB$. When the power of the target is low, only GLRT detector can be used in order to insure low probability of false alarm with sufficient probability of detection.
Figures 3.23, 3.24 and 3.26 represent the probability of detection with respect to the normalized frequency of the simulated target for a Probability of False Alarm (or PFA), i.e. the probability to wrongly detect a target, fixed to 10^{-2}. Robust estimators are competitive with respect to the non robust estimators (which have better performance since there is no outliers or misspecification). Only ACG normalized estimator show huge weaknesses.

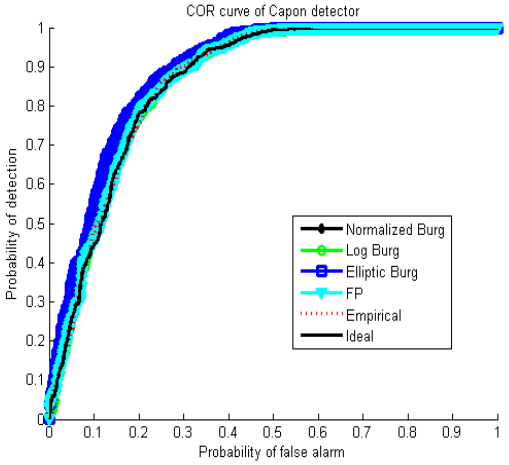
Remark 3.51. *The chosen PFA (set to* 10^{-2}*) is too high for realistic radar applications. This choice is due to the time of computation of the estimators and our limited computer resources. It gives nevertheless a realistic intuition on the hierarchy between the different presented estimators.*



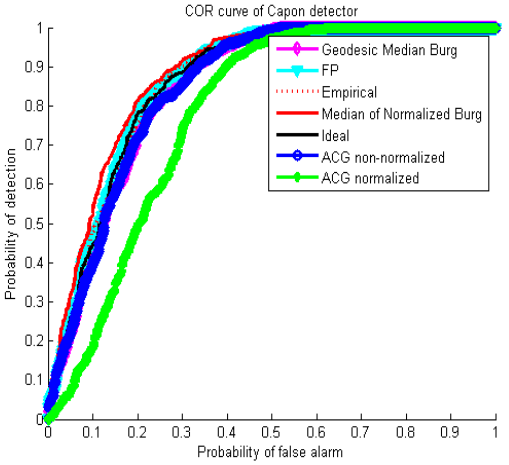
(a) GLRT detector, non robust estimators



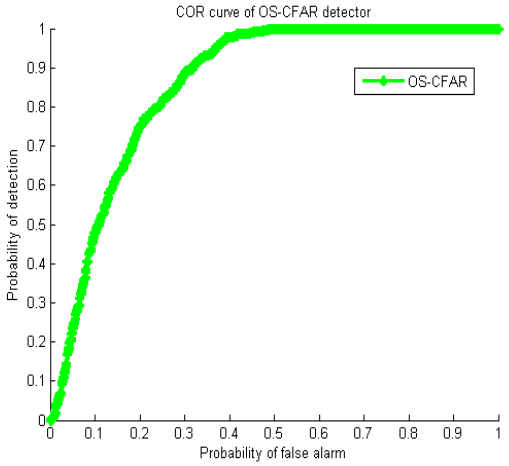
(b) GLRT detector, robust estimators



(c) Capon detector, non robust estimators

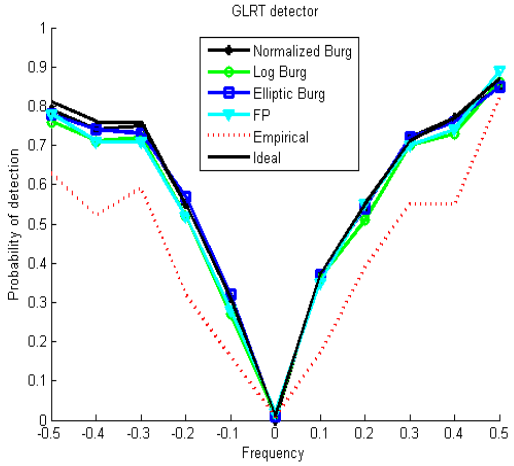


(d) Capon detector, robust estimators

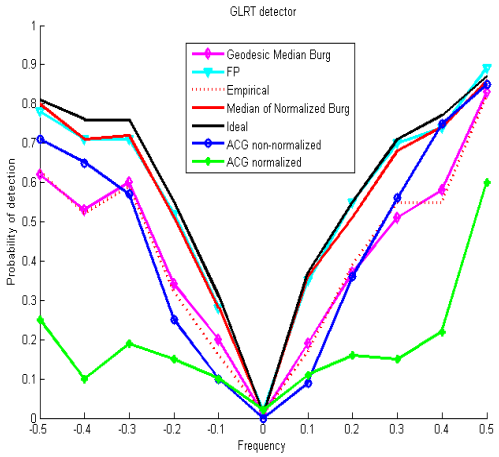


(e) OS-CFAR detector

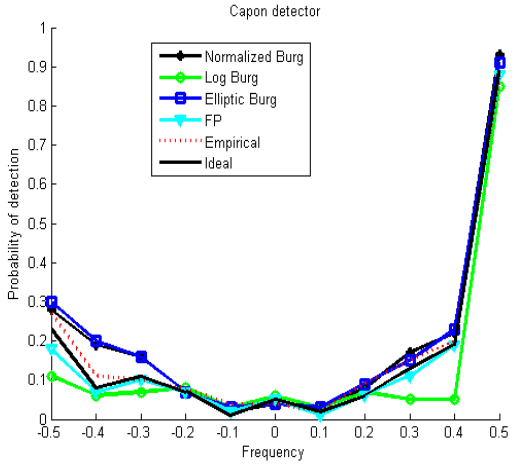
Figure 3.22 – COR curves for $\nu = 0.6$, a clutter level set to $20dB$ for a target of power $20dB$ (Scenario 0) : the probability of detection is averaged among all possible normalized frequencies of a target (the scale of the PFA is logarithmic)



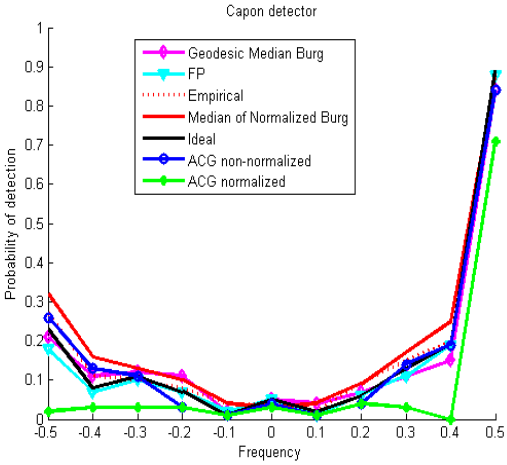
(a) GLRT detector, non robust estimators



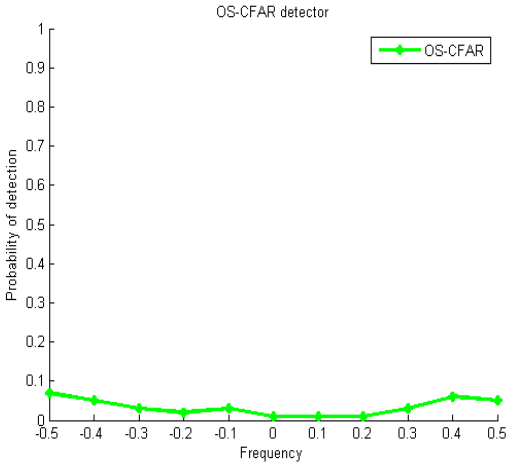
(b) GLRT detector, robust estimators



(c) Capon detector, non robust estimators

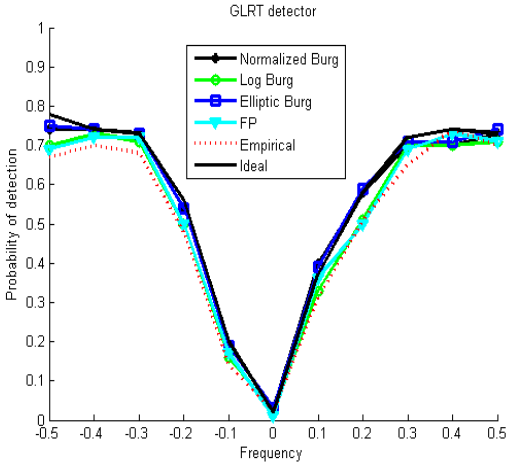


(d) Capon detector, robust estimators

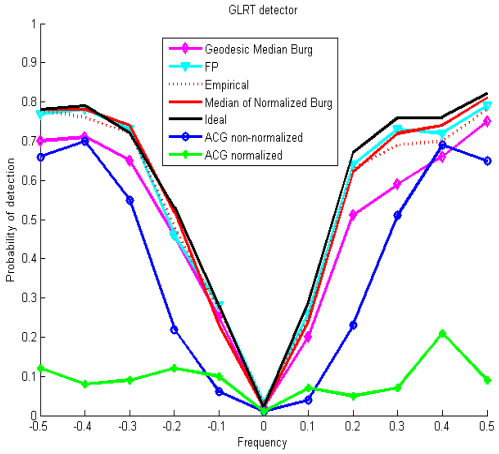


(e) OS-CFAR detector

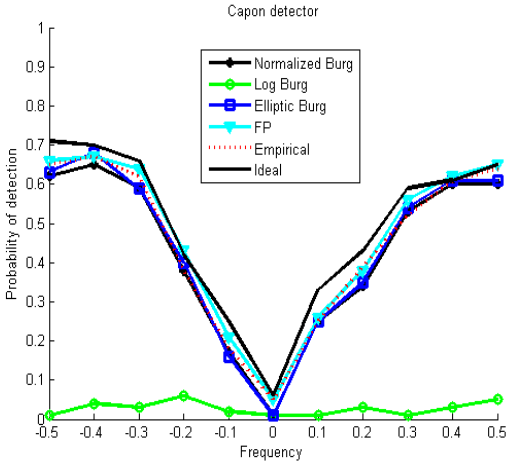
FIGURE 3.23 – Probability of detection in function of the frequency of the target for $\nu = 0.6$, a clutter level set to $20dB$ for a target of power $20dB$, a desired PFA equal to 10^{-2} (Scenario 0)



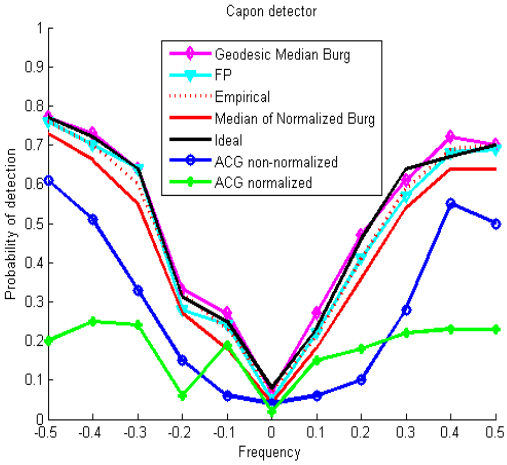
(a) GLRT detector, non robust estimators



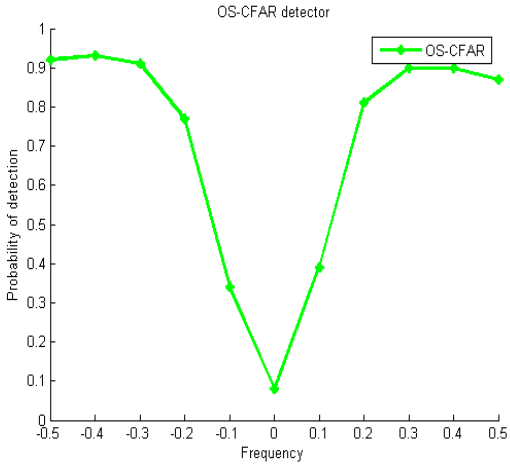
(b) GLRT detector, robust estimators



(c) Capon detector, non robust estimators

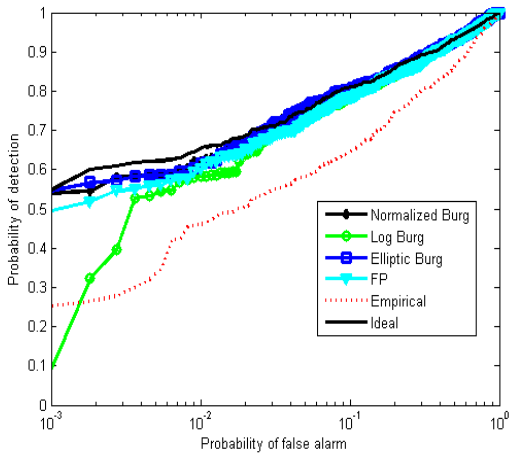


(d) Capon detector, robust estimators

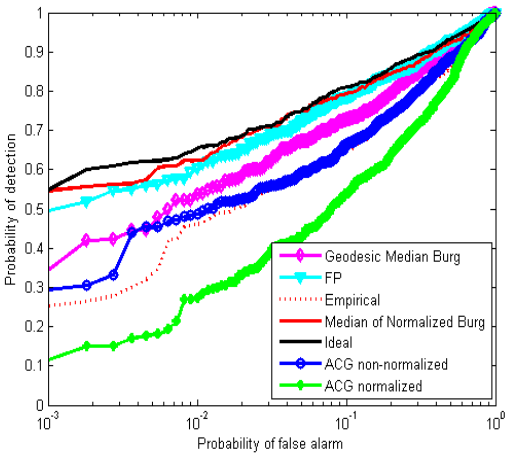


(e) OS-CFAR detector

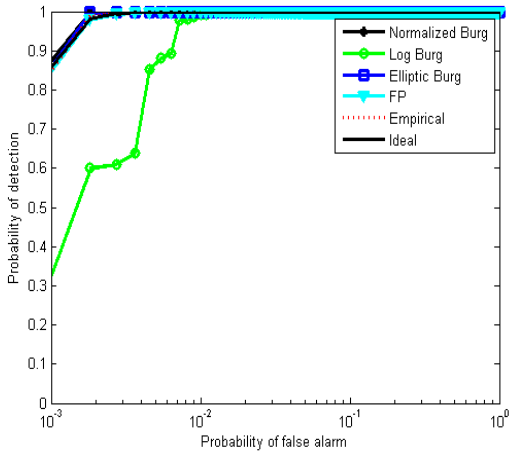
Figure 3.24 – Probability of detection in function of the frequency of the target for $\nu = 3$, a clutter level set to $20dB$ for a target of power $20dB$, a desired PFA equal to 10^{-2} (Scenario 0)



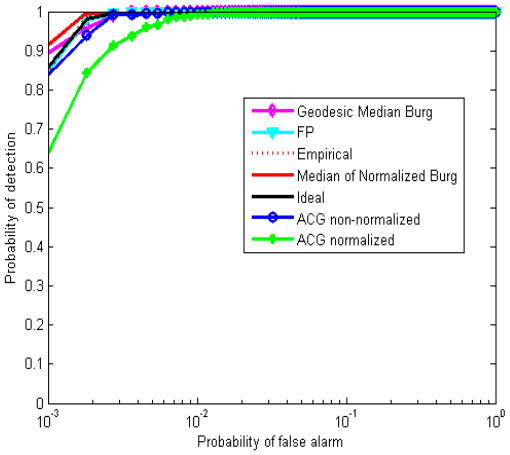
(a) GLRT detector, non robust estimators



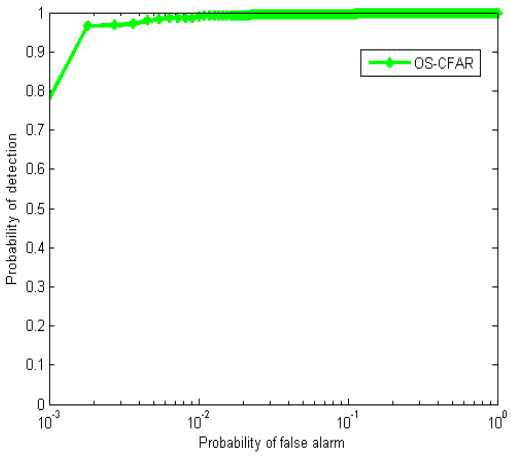
(b) GLRT detector, robust estimators



(c) Capon detector, non robust estimators

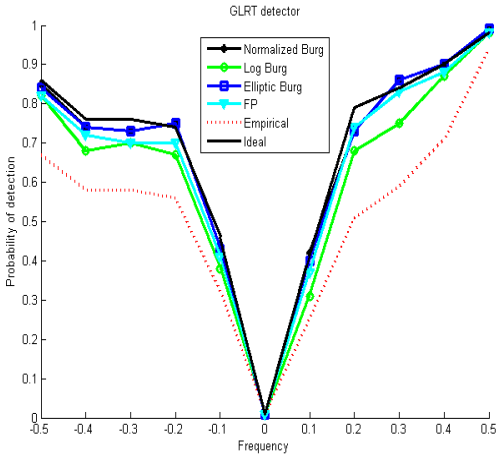


(d) Capon detector, robust estimators

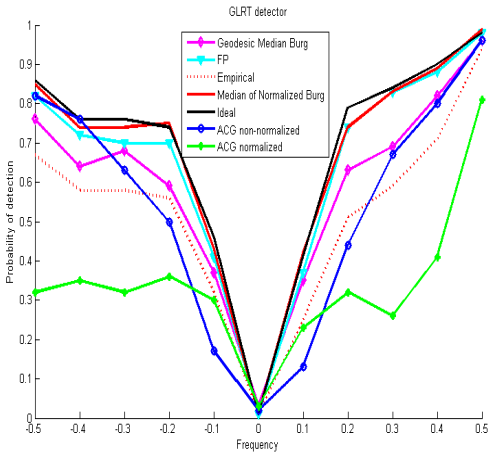


(e) OS-CFAR detector

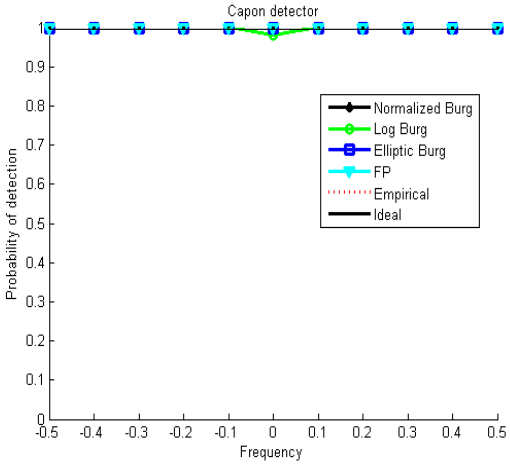
FIGURE 3.25 – COR curves for $\nu = 0.6$, a clutter level set to $20dB$ for a target of power $40dB$ (Scenario 0) : the probability of detection is averaged among all possible normalized frequencies of a target (the scale of the PFA is logarithmic)



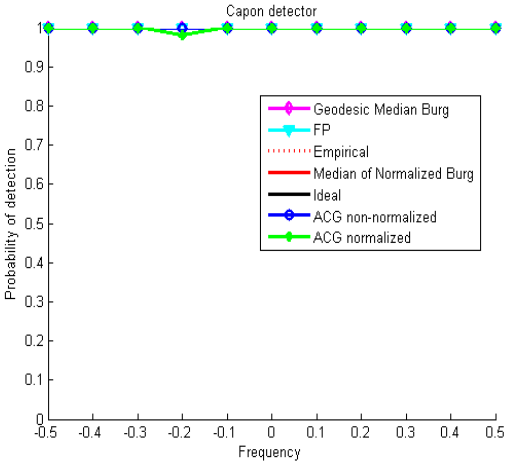
(a) GLRT detector, non robust estimators



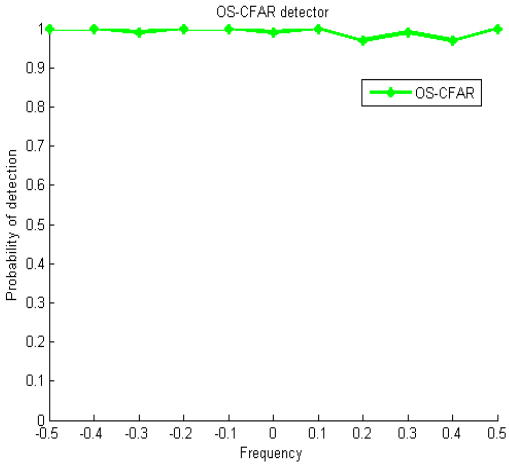
(b) GLRT detector, robust estimators



(c) Capon detector, non robust estimators



(d) Capon detector, robust estimators



(e) OS-CFAR detector

Figure 3.26 – Probability of detection in function of the frequency of the target for $\nu = 0.6$, a clutter level set to $20dB$ for a target of power $40dB$, a desired PFA equal to 10^{-2} (Scenario 0)

3.9.3 Scenario 1 : multiple targets

Scenario 1 illustrates the robustness of algorithms with respect to a light contamination. We pollute the ambiance samples $z_1, ..., z_N$ of clutter noise by 10 target samples of power set to $40dB$ or $20dB$ and of normalized frequency set to 0.3 (see Figure 3.27). This allow us to simulate the influence of an other target such as an airplane in the ambiance cases. Figures 3.28, 3.29 and 3.30 illustrate the probability of detection with respect to the normalized frequency of the introduced target for a PFA set to 10^{-2}. The principal difference with the scenario 0 is the behavior of each estimator for a target of frequency 0.3, i.e. close to the frequency of contaminating targets. We see that the so-called "Geodesic Median Burg", "ACG Normalized Burg" and "Median of Normalized Burg" estimators are not disturbed by the contamination, contrary to all other estimators.

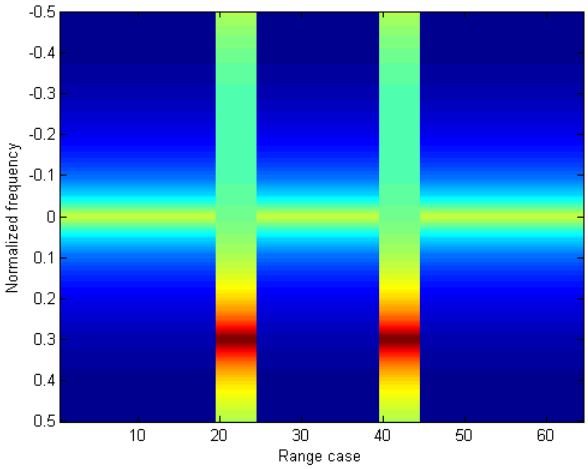
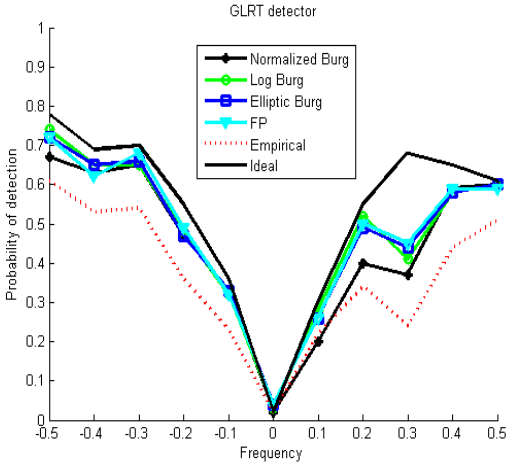
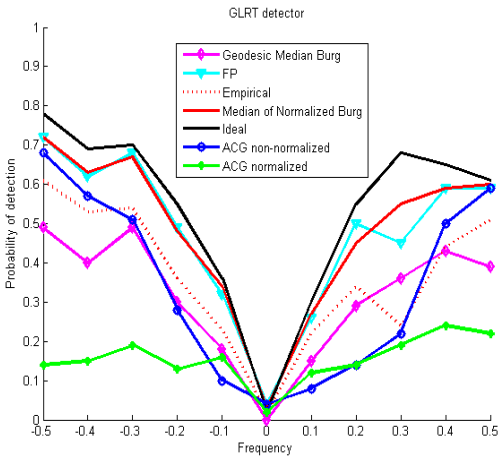


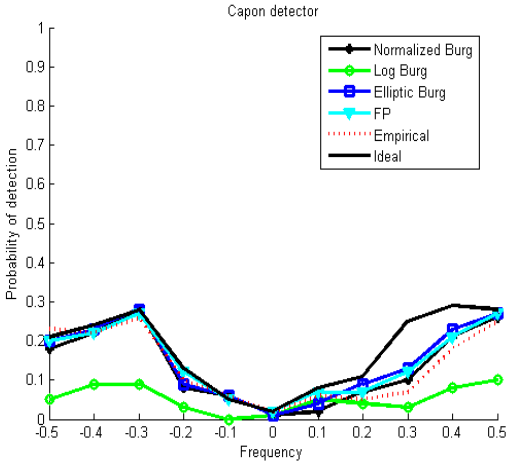
FIGURE 3.27 – Simulated spectra of the surrounding range cases : an ambient clutter of power 20dB and frequency 0 and two contaminating targets (corresponding to 5 samples each) of power 40dB and frequency 0.3



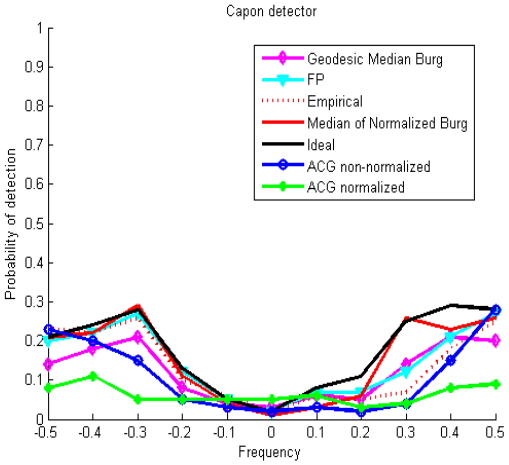
(a) GLRT detector, non robust estimators



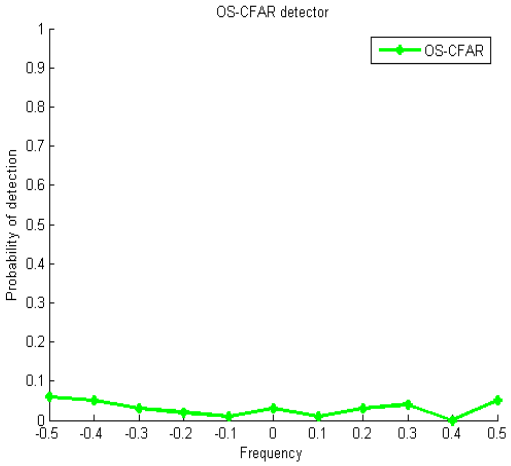
(b) GLRT detector, robust estimators



(c) Capon detector, non robust estimators

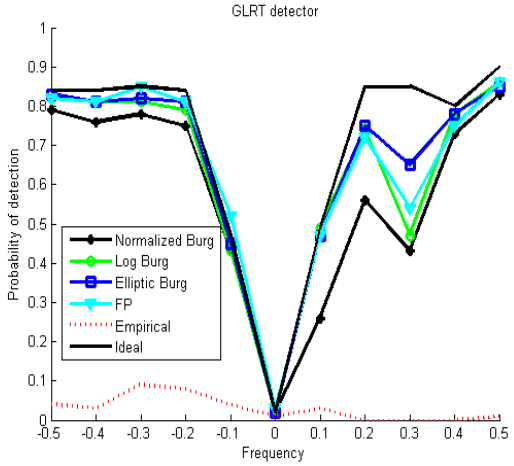


(d) Capon detector, robust estimators

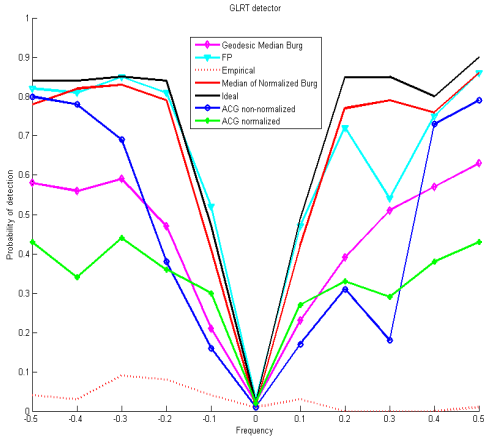


(e) OS-CFAR detector

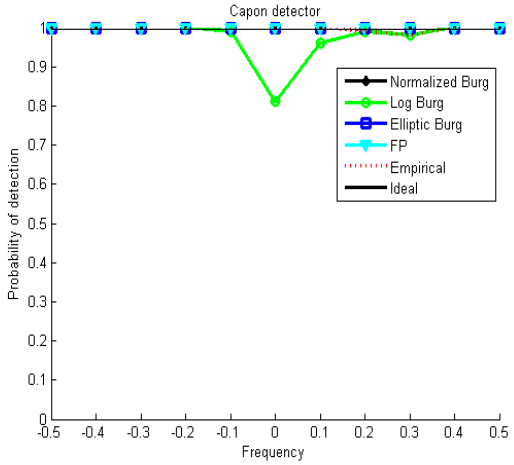
Figure 3.28 – Probability of detection in function of the frequency of the target for $\nu = 0.6$, a clutter level set to $20dB$ for a target of power $20dB$, contaminating targets with power $20dB$, a desired PFA equal to 10^{-2} (Scenario 1)



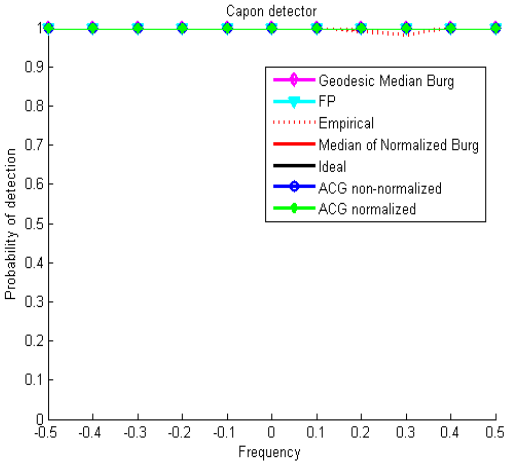
(a) GLRT detector, non robust estimators



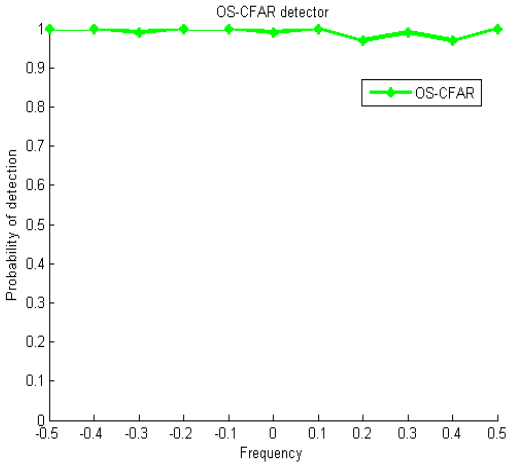
(b) GLRT detector, robust estimators



(c) Capon detector, non robust estimators

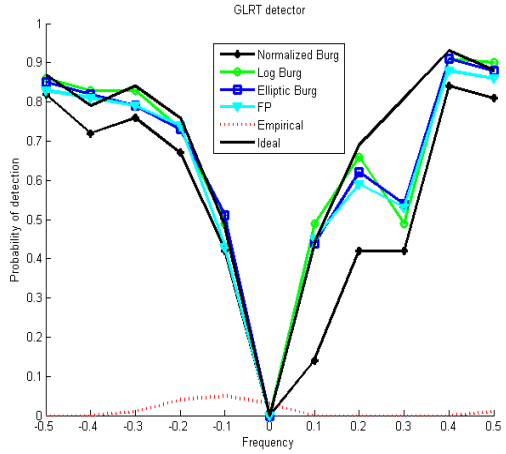


(d) Capon detector, robust estimators

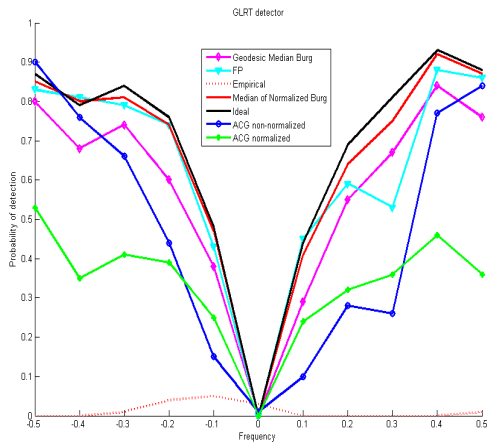


(e) OS-CFAR detector

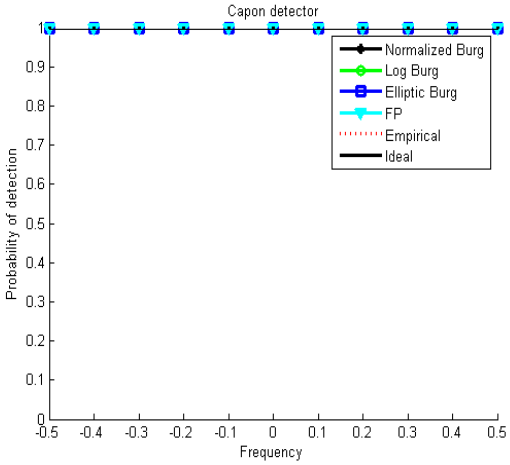
FIGURE 3.29 – Probability of detection in function of the frequency of the target for $\nu = 0.6$, a clutter level set to $20dB$ for a target of power $40dB$, contaminating targets with power $40dB$, a desired PFA equal to 10^{-2} (Scenario 1)



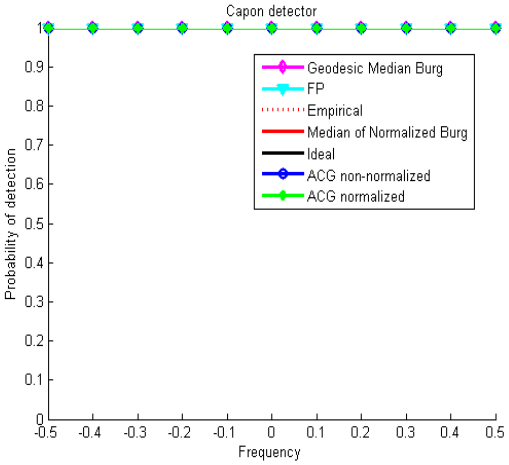
(a) GLRT detector, non robust estimators



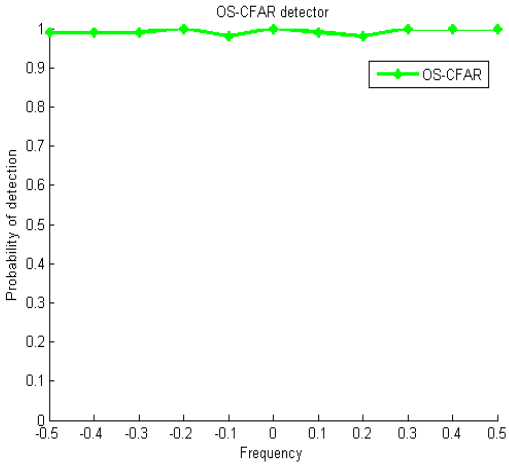
(b) GLRT detector, robust estimators



(c) Capon detector, non robust estimators



(d) Capon detector, robust estimators



(e) OS-CFAR detector

Figure 3.30 – Probability of detection in function of the frequency of the target for $\nu = 3$, a clutter level set to $20dB$ for a target of power $40dB$, contaminating targets with power $40dB$, a desired PFA equal to 10^{-2} (Scenario 1)

3.9.4 Scenario 2 : clutter transition

Finally, we present a Scenario of a clutter transition between a clutter of frequency 0 (that could represent a ground clutter) and a clutter of frequency 0.3 (that could represent a sea clutter). The power of the clutters is set to $20dB$ (Figure 3.31). Figure 3.33 shows the estimated spectra for the 64 ambient cells around each cell under test. The result is then to compare with the true spectra in Figure 3.31. We see that only "Geodesic Median Burg", "ACG Normalized Burg" and "Median of Normalized Burg" correctly estimate the transition. These estimators are said to be robust to a large contamination of the data. The same information is illustrated by Figure 3.32. Indeed, we show the value of the GLRT detector for the cell under test where a target of power $20dB$ has been introduced at the frequency of the second clutter (i.e. 0.3). We see that most estimators could not detect the target anymore in the zone close to the transition. This situation could correspond to the case of a pop-up target which would "hide" in the clutter.

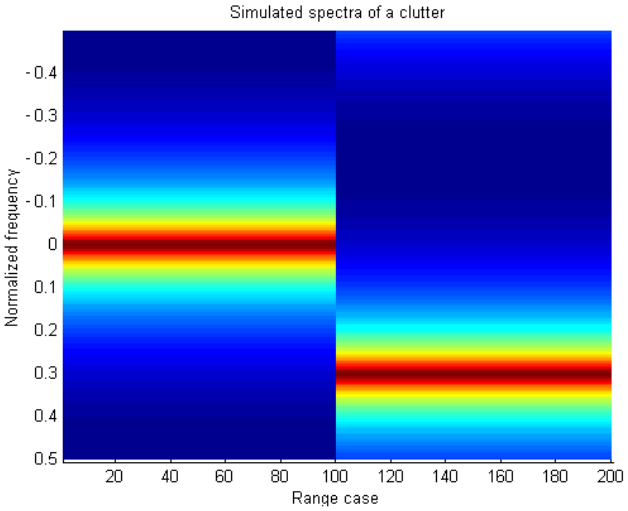


FIGURE 3.31 – Spectra of the simulated scenario of change of clutter (modeled by two $AR(1)$)

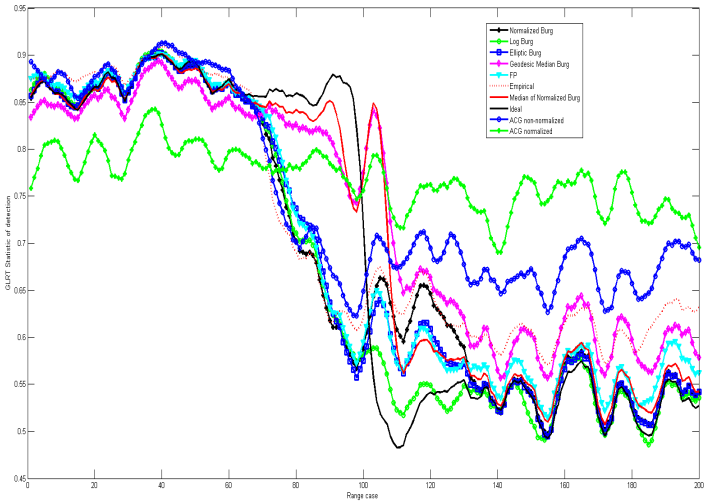
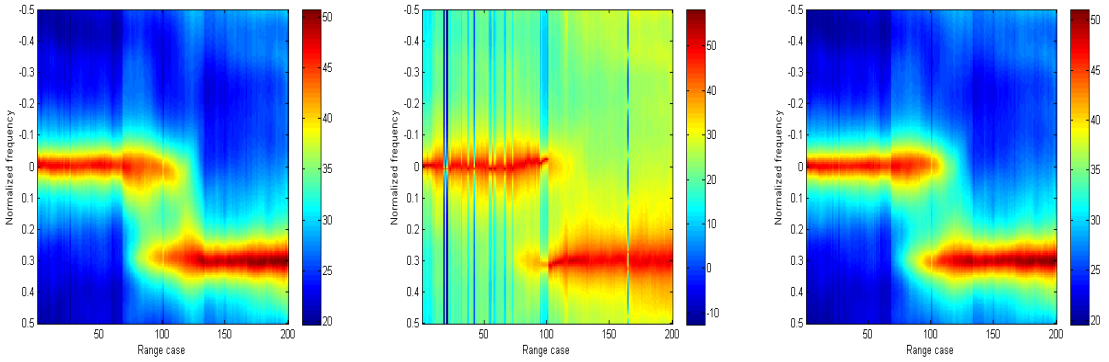
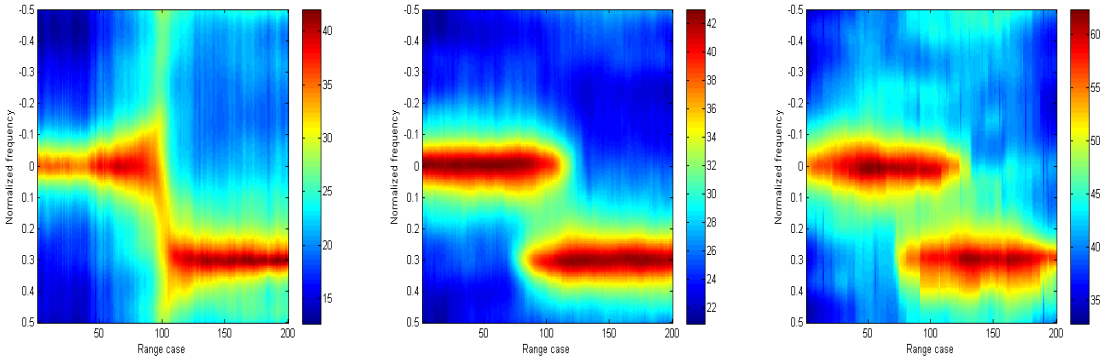


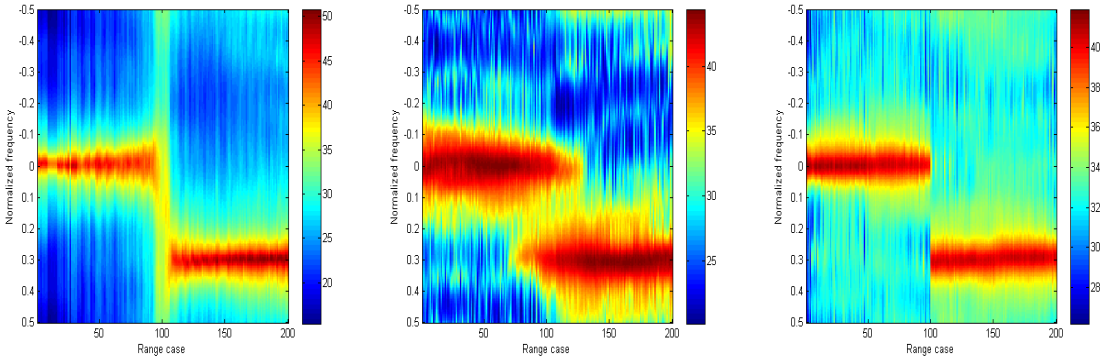
Figure 3.32 – GLRT detector mean value under H_1 with a target at normalized frequency 0.3



(a) Normalized Burg estimator (b) Log Burg estimator (c) Elliptical Burg estimator



(d) Geodesic Median Burg estimator (e) Fixed Point estimator (f) Empirical estimator



(g) Median of Normalized Burg estimator (h) ACG non-normalized estimator (i) ACG normalized estimator

FIGURE 3.33 – Estimated spectra for the scenario 2 illustrated by Figure 3.31 of different estimators

3.9.5 Computation time

For illustrative purposes, Table 3.34 gives the mean run time for one estimation of a covariance for each algorithm with a QuadriCore processor on a Matlab implementation.

Algorithm	Mean execution time for N = 32 (in milliseconds)	Mean execution time for N = 64 (in milliseconds)
Empirical covariance	0.2	0.4
Fixed Point	6.8	8.7
Geodesic Median Burg ($M = 7$)	16.8	31.8
Geodesic Median Burg ($M = 3$)	13.8	25.2
Normalized Burg ($M = 7$)	1.7	2.2
Normalized Burg ($M = 3$)	0.9	1.1
Log Burg ($M = 7$)	10.6	14.3
Log Burg ($M = 3$)	4.8	6.0
Elliptical Burg ($M = 7$)	6.5	8.6
Elliptical Burg ($M = 3$)	3.8	3.5
Median of Normalized Burg ($M = 7$)	8.7	9.5
Median of Normalized Burg ($M = 3$)	3.7	3.6
ACG non-normalized	1773.6	1860.1
ACG normalized	3140.6	3285.4

Figure 3.34 – Mean execution time of different algorithm (M represents the order of the autoregressive model)

3.9.6 Simulations analysis

The presented simulations show a difference of behavior of the introduced estimators with respect to their robustness.
First, we clearly observe the difference between the GLRT and the Capon detector. The higher the SCR (Signal to Clutter Ratio) is, the more efficient the Capon detector with respect to GLRT is. Indeed, contrary to GLRT detector, Capon's detector takes into account the difference of power between the cell under test and the ambiance. A thresholding of the SCR seems to be inevitable before choosing one detector.
Secondly, the introduced ACG Non-Normalized and ACG Normalized estimators lack efficiency. It can be seen on performances of those estimators illustrated by Figure 3.20 but also on figures representing the probability of detection with respect to normalized frequency. However, the robustness of these procedures is clearly illustrated by Figures 3.33 and 3.32. Especially, ACG Normalized estimator perfectly estimate the change of clutter on Figure 3.33.
Among non robust estimators, the so-called Normalized Burg and Elliptic Burg seem to have similar behavior. We often encounter numerical problems of convergence of Log Burg estimator. This could explain the degradation of its performances with respect to the latter Burg estimators.
The good news is that the robust versions of Burg estimators, namely "Geodesic Median Burg" and "Median of Normalized Burg" estimators, show almost no lack of efficiency as ACG robust estimators. They furthermore show robustness properties as expected in the two scenarios (illustrated by Figures 3.23-3.26 and Figures 3.28-3.30). Under targets contamination (Scenario 1), they have comparable performances as the reference estimator – namely the FP (or Tyler's) estimator. Under heavy contamination (Scenario 2), they both show a robustness with respect to the clutter transition (Figures 3.33-3.30) contrary to the FP estimator.
Among the two geodesic Burg estimators, good performances of the Median of Normalized Burg estimators should be tempered by the following remark : the estimation on 8 sub-

samples is well adapted to our simulation scenarios since each subsample will mostly be homogeneously distributed. This would not be the case if the sample were been drawn randomly from one of the two clutter distribution. In other words, the Median of Normalized is particularly well adapted to clutter transitions.

3.10 Appendix

3.10.1 Proof of the asymptotic bias of Log-Burg estimator

Let us define the function for $z = (z_1, z_2) \in \mathbb{C}^2$

$$u(z, \mu) = 2 \log \Big((1 + |\mu|^2)(|z_1|^2 + |z_2|^2) + 2\Re(\mu z_1 \overline{z_2}) \Big)$$

We will denote by μ_T the true parameter of the law of each couple $z_{i,n} = (f_{i,m}(n), b_{i,m}(n-1)) \in \mathbb{C}^2$. The estimator is then the minimizer :

$$\hat{\mu}_m = \arg \min_{|\mu|<1} \sum_{i=1}^{N} \sum_{n=m+1}^{d} u(z_{i,n}, \mu).$$

By assumption, this minimizer is unique and is then determined by the equation :

$$\psi_n(\mu) := \frac{1}{N(d-m-1)} \sum_{i=1}^{N} \sum_{n=m+1}^{d} \nabla_\mu u(z_{i,n}, \mu) = 0$$

As for all μ such that $|\mu| < 1$ and $z = (z_1, z_2) \in \mathbb{C}$, $|\nabla_\mu u(z, \mu)| \leq \frac{2|z_1|^2 + 2|z_2|^2 + 4|z_1||z_2|}{|z_1|^2 + |z_2|^2}$, the uniform law of large numbers allows to assert :

$$\sup_{|\mu|<1} |\psi_n(\mu) - \mathbb{E}[\nabla_\mu u(Z, \mu)]| \to 0 \text{ a.s.}$$

Then, if we note $\psi(\mu) = \mathbb{E}[\nabla_\mu u(Z, \mu)]$, we have :

$$|\psi(\hat{\mu}_m)| = |\psi(\hat{\mu}_m) - \psi_n(\hat{\mu}_m)| \to 0 \text{ a.s.}$$

It remains for us to study the function ψ and find its unique zero in order to conclude. Let us recall that if $Z = (Z_1, Z_2)$ is a Gaussian vector of mean zero and covariance $\Sigma = \begin{pmatrix} 1 & -\mu_T \\ -\bar{\mu}_T & 1 \end{pmatrix}$:

$$\psi(\mu) = \mathbb{E}\left[\frac{\mu \|Z\|^2 + 2 Z_1 \overline{Z_2}}{(1 + |\mu|^2)\|Z\|^2 + 2\Re(\mu Z_2 \overline{Z_1})} \right]$$

Let us define the matrices :

$$M = \begin{pmatrix} 1 & \mu \\ \bar{\mu} & 1 \end{pmatrix}$$

and

$$L = \mathbb{E}\left[\frac{Z Z^+}{Z^+ M^2 Z} \right]$$

Then $\psi(\mu) = 2L_{12} + \mu(L_{11} + L_{22})$. Let us compute L for $\mu = y \frac{\mu_T}{|\mu_T|}$ with $y > 0$. The singular value decomposition of M and Σ are then

$$\Sigma = UEU^+ \, ; \, M = UDU^+$$

with

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & \frac{\mu_T}{|\mu_T|} \\ -\frac{\mu_T}{|\mu_T|} & 1 \end{pmatrix}$$

and

$$E = \begin{pmatrix} 1+|\mu_T| & 0 \\ 0 & 1-|\mu_T| \end{pmatrix} ; D = \begin{pmatrix} 1+y & 0 \\ 0 & 1-y \end{pmatrix}$$

Denote by X a Gaussian vector of covariance $M\Sigma M$. We express L in function of X

$$L = M^{-1}\mathbb{E}\left[\frac{XX^+}{X^+X}\right]M^{-1}$$

$\mathbb{E}\left[\frac{XX^+}{X^+X}\right]$ has been computed by Bausson et al. in [16]. If $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = DED^+$ and $\Delta = \begin{pmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{pmatrix}$,

$$\mathbb{E}\left[\frac{XX^+}{X^+X}\right] = U\Delta U^+$$

with

$$\begin{cases} \lambda_1 = \frac{(1+y)^2}{1+|\mu_T|} \\ \lambda_2 = \frac{(1-y)^2}{1-|\mu_T|} \end{cases}$$

and

$$\begin{cases} \delta_1 = \frac{\lambda_1\lambda_2}{\lambda_2-\lambda_1}\left(\frac{\log(\lambda_2)-\log(\lambda_1)}{\lambda_2-\lambda_1} - \frac{1}{\lambda_2}\right) \\ \delta_2 = \frac{\lambda_1\lambda_2}{\lambda_1-\lambda_2}\left(\frac{\log(\lambda_2)-\log(\lambda_1)}{\lambda_2-\lambda_1} - \frac{1}{\lambda_1}\right) \end{cases}$$

Then $L = UD^{-1}\Delta D^{-1}U^+$. We finally can give an explicit expression of ψ :

$$\psi(\mu) = \psi(\mu) = \frac{\mu}{|\mu|}\left(\frac{\delta_1}{1+y} - \frac{\delta_2}{1-y}\right)$$

We can now search the zero of this function which depends only on $y = |\mu|$ and $|\mu_T|$:

$$\begin{aligned} \psi(\mu) = 0 &\Leftrightarrow \frac{\delta_1}{1+y} = \frac{\delta_2}{1-y} \\ &\Leftrightarrow \frac{\log(\lambda_2)-\log(\lambda_1)}{\lambda_2-\lambda_1}\left(1+y+\frac{1}{1-y}\right) = \frac{1}{\lambda_2(1+y)} + \frac{1}{\lambda_1(1-y)} \\ &\Leftrightarrow 2\frac{\log(\lambda_2/\lambda_1)}{\lambda_2/\lambda_1-1} = \frac{1-y}{\lambda_2/\lambda_1} + 1 + y \\ &\Leftrightarrow y = \frac{2a\log(a)}{(1-a)^2} + \frac{1+a}{1-a} \text{ and } a = \lambda_2/\lambda_1 = \frac{(1-y)^2(1+|\mu_T|)}{(1+y)^2(1-|\mu_T|)} \end{aligned}$$

If we note $B_2 : [0;1] \to [0;1], a \mapsto \frac{2a\log(a)}{(1-a)^2} + \frac{1+a}{1-a}$, B_2 is invertible and the y solution to :

$$\frac{(1-y)^2(1+|\mu_T|)}{(1+y)^2(1-|\mu_T|)} = B_2^{-1}(y)$$

i.e. $|\mu_T| = g(y)$ with

$$g(y) = \frac{(1+y)^2B_2^{-1}(y) - (1-y)^2}{(1+y)^2B_2^{-1}(y) + (1-y)^2}$$

We can conclude that the solution of $\psi(\mu) = 0$ is unique and as ψ is defined on a compact $\{\mu \in \mathbb{C}, |\mu| \leq 1\}$ and $\psi(\hat{\mu}_m) \to 0$ a.s., we have :

$$\hat{\mu}_m \to g^{-1}(|\mu_T|)\frac{\mu_T}{|\mu_T|} \text{ a.s.}$$

Bibliographie

[1] P-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[2] S. M. Ali and S.D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28(1) :131–142, 1966.

[3] S. Amari and A. Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences, Technical Sciences*, 58(1) :183–195, 2010.

[4] L. Ambrosio, N. Gigli, and G. Savare. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zurich. Birkhauser, Verlag, Basel, 2005.

[5] R.D. Anderson and V. L. Klee Jr. Convex functions and upper semicontinuous collections. *Duke Math. Journal*, 19(2) :349–357, 1952.

[6] M. Arnaudon, F. Barbaresco, and L. Yang. Riemannian medians and means with applications to radar signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 7(4) :595–604, 2013.

[7] M. Arnaudon and L. Miclo. A stochastic algorithm finding p-means on the circle. *arXiv :1301.7156*, 2013.

[8] F. Aurenhammer. Power diagrams : properties, algorithms and applications. *SIAM Journal on Computing*, 16(1) :78–96, 1987.

[9] F. Aurenhammer, F. Hoffmann, and B. Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1) :61–76, 1998.

[10] F. Barbaresco. Super resolution spectrum analysis regularization : Burg, capon and ago antagonistic algorithms. *Proc. EUSIPCO'96, Trieste*, pages 2005–2008, 1996.

[11] F. Barbaresco. New foundation of radar doppler signal processing based on advanced differential geometry of symmetric spaces : Doppler matrix cfar and radar application. *Radar09 Conference, Bordeaux*, 2009.

[12] F. Barbaresco. Information geometry of covariance matrix : Cartan-siegel homogeneous bounded domains, mostow/berger fibration and frechet median. *Matrix Information Geometry*, 2012.

[13] F. Barbaresco. Koszul information geometry and souriau geometric temperature/capacity of lie group thermodynamics. *Entropy*, 16 :4521–4565, 2014.

[14] T.J. Barnard and D.D. Weiner. Non-gaussian clutter modeling with generalized spherically invariant random vectors. *IEEE Trans.-SP*, 44(10) :2384–2390, 1996.

[15] B.A. Basu, I.R. Harris, N.L. Hjort, and M.C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3) :549–559, 1998.

[16] S. Bausson, F. Pascal, P. Forster, J.P. Ovarlez, and P. Larzabal. First- and second-order moments of the normalized sample covariance matrix of spherically invariant random vectors. *IEEE Signal Processing Letters*, 14(6) :425–428, 2007.

[17] J. Bensadon. Black-box optimization using geodesics in statistical manifolds. *arXiv :1309.7168v2*, 2013.

[18] M. Berkane, K. Oden, and P.M. Bentler. Geodesic estimation in elliptical distributions. *Journal of Multivariate Analysis*, 63(1) :35–46, 1997.

[19] P. Bertail. Empirical likelihood in some semiparametric models. *Bernoulli*, 12(2) :299–331, 2006.

[20] O. Besson and Y. Abramovich. On the fisher information matrix for multivariate elliptically contoured distributions. *IEEE Signal Processing Letters*, 20(11) :1130–1133, 2013.

[21] R. Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.

[22] D.A. Bini, B. Iannazzo, B. Jeuris, and R. Vandebril. Geometric means of structured matrices. *BIT Numerical Mathematics*, 54(1) :55–83, 2014.

[23] X. Blanc, C. Le Bris, and P.-L. Lions. From molecular models to continuum mechanics. *Archive for Rational Mechanics and Analysis*, 164(4) :341–381, 2002.

[24] X. Blanc, C. Le Bris, and P.-L. Lions. Atomistic to continuum limits for computational materials science. *Mathematical Modelling and Numerical Analysis*, 41(2) :391–426, 2007.

[25] J.M. Borwein and A.S. Lewis. Duality relationships for entropy-like minimization problems. *SIAM Journal of Control and Optimization*, 29 :325–338, 1991.

[26] G.E.P. Box and D.R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2) :211–252, 1964.

[27] Y. Brenier. Polar factorization and monotone rearrangement of vector valued functions. *Communications on Pure and Applied Mathematics*, 44(4) :375–417, 1991.

[28] P.J. Brockwell and R. Dalhaus. Generalized durbin-levinson and burg algorithms. *Journal of Econometrics*, 118(1-2) :129–149, 2003.

[29] M. Broniatowski and A. Keziou. Divergences and duality for estimation and test under moment condition models. *Journal of Statistical Planning and Inference*, 142(9) :2554–2573, 2012.

[30] M. Broniatowski and I. Vajda. Several applications of divergence criteria in continuous families. *Kybernetika*, 48(4) :600–636, 2012.

[31] J.P. Burg. Maximum entropy spectral analysis. *Modern Spectrum Analysis, D.G. Childers, ed., IEEE Press, New York*, pages 34–41, 1978.

[32] S. Cambanis, S. Huang, and G. Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3) :368–385, 1981.

[33] J. Capon. High resolution frequency-wavenumber spectral analysis. *Proceedings of the IEEE*, 57(8) :1408 –1418, 1969.

[34] G. Carlier, A. Galichon, and F. Santambrogio. From knothe's transport to brenier's map and a continuation method for optimal transport. *SIAM Journal of Mathematical Analysis*, 41(6) :2554–2576, 2009.

[35] J. Carretero-Moya, J. Gismero-Menoyo, A. Asensio-Lopez, and A. Blanco-Del-Campo. Small-target detection in high-resolution heterogeneous sea-clutter : An empirical analysis. *IEEE Transactions on Aerospace and Electronic Systems*, 47(3) :1880–1898, 2011.

[36] E. Cartan. Groupes simples clos et ouverts et géométrie riemannienne. *J. Math. pures et appl.*, 8 :1–33, 1929.

[37] L.K. Chan. On a characterization of distributions by expected values of extreme order statistics. *Amer.Math.Monthly*, 74 :950–951, 1967.

[38] J.T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3) :287–317, 1997.

[39] P. Chaudhuri. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434) :862–872, 1996.

[40] E. Conte, A. De Maio, and G. Galdi. Recursive estimation of the covariance matrix of a compound-gaussian process and its application to adaptative cfar detection. *IEEE Transactions on Signal Processing*, 50(8) :1908–1915, 2002.

[41] E. Conte and M. Longo. Characterization of radar clutter as a spherically invariant random process. *IEE Proc.-Pt.F*, 134(2) :191 –197, 1987.

[42] I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutato Int. Kozl*, 8 :85–108, 1963.

[43] I. Csiszár, F. Gamboa, and E. Gassiat. Mem pixel correlated solutions for generalized moment and interpolation problems. *IEEE Trans. Inform. Theory*, 45(7) :2253–2270, 1999.

[44] I. Csiszár and F. Matúš. Generalized minimizers of convex functionals, bregman distance, pythagorean identities. *Kybernetika*, 48(4) :637–689, 2012.

[45] H.A. David. *Order Statistics*. 2nd edition. New York, Wiley, 1981.

[46] M.P. do Carmo. *Riemannian geometry*. 1st Edition. Birkhauser, 1992.

[47] G.S. Easton and R.E. McCulloch. A multivariate generalization of quantile-quantile plots. *Journal of the American Statistical Association*, 85(410) :376–386, 1990.

[48] S. Eguchi and Y. Kano. Robusting maximum likelihood estimation by psidivergence. *(ISM Research Memorandum 802), Tokyo : Institute of Statistical Mathematics*, 2001.

[49] J.H.J. Einmahl and D.M. Mason. Generalized quantile process. *Annals of Statistics*, 20(2) :1062–1078, 1992.

[50] E. Elamir and A. Seheult. Trimmed l-moments. *Computational Statistics and Data Analysis*, 43(3) :299–314, 2003.

[51] M. Fréchet. Les éléments aléatoires de natures quelconque dans un espace distancié. *Annales de l'IHP*, 10(4) :215–310, 1948.

[52] H. Fujisawa. Normalized estimating equation for robust parameter estimation. *Electronic Journal of Statistics*, 7 :1587–1606, 2013.

[53] H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9) :2053–2081, 2008.

[54] A. Galichon and M. Henry. Trimmed l-moments. *Journal of economic theory*, 147(4) :1501–1516, 2012.

[55] F. Gini. Sub-optimum coherent radar detection in a mixture of k-distributed and gaussian clutter. *Proc. Inst. Electr. Eng.*, 144(1) :39–48, 1997.

[56] V.P. Godambe and M.E. Thompson. An extension of quasi-likelihood estimation. *Journal of Statistical Planning and Inference*, 22(2) :137–172, 1989.

[57] C. Gourieroux and J. Jasiak. Dynamic quantile models. *Journal of econometrics*, 147(1) :198–205, 2008.

[58] X. Gu, F. Luo, J. Sun, and S.-T. Yau. Variational principles for minkowski type problems, discrete optimal transport, and discrete monge-ampere equations. *arXiv :1302.5472*, 2013.

[59] M. Hallin, H. Oja, and D. Paindaveine. Semiparametrically efficient rank-based inference for shape ii. r-estimation of shape. *Annals of Statistics*, 34(6) :2757–2789, 2006.

[60] F.R. Hampel, E.M. Ronchetti, P.J. Rousseuw, and WA.A Stahel. *Robust Statistics : The approach Based on Influence Functions*. New York : Willey, 1986.

[61] L.P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4) :1029–1054, 1982.

[62] L.P. Hansen. Finite-sample properties of some alternative gmm estimators. *Journal of Business and Economic Statistics*, 14(3) :262–280, 1996.

[63] J.R. Hosking. Some theoretical results concerning l-moments. *Research report RC14492*, 1989.

[64] J.R. Hosking. L-moments : analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society*, 52(1) :105–124, 1990.

[65] J.R. Hosking and J.R. Wallis. *Regional Frequency Analysis : An approach based on L-moments*. Cambridge University Press, 1997.

[66] P.J. Huber and E.M. Ronchetti. *Robust Statistics*. Second Edition. John Wiley and Sons Inc., 2009.

[67] E. Jakeman. On the statistics of k-distributed noise. *Journal of Physics A : Mathematics and General*, 13(1) :31–48, 1980.

[68] E. Jurczenko, B. Maillet, and P. Merlin. Hedge fund portfolio selection with higher-order moments : a nonparametric mean-variance-skewness-kurtosis efficient frontier. *Multi-moment asset allocation and pricing models*, pages 21–66, 2006.

[69] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied Mathematics*, 30(5) :509–541, 1977.

[70] H. Karcher. Riemannian center of mass and so called karcher mean. *arXiv :1407.2087v1*, 2014.

[71] A.G. Konheim. A note on order statistics. *Amer.Math.Mon*, 78 :524, 1971.

[72] S. Kraut, L.L. Scharf, and L.T. McWhorter. Adaptative subspace detectors. *IEEE Transactions on Signal Processing*, 49(1) :1–16, 2001.

[73] K. Kreutz-Delgado. The complex gradient operator and the $\mathbb{CR}$ calculus. *http ://arxiv.org/abs/0906.4835v1*, 2009.

[74] C. Le Bris. *Systèmes multi-échelles : modélisation et simulation*. SMAI, Mathématiques et Applications, 47. 2005.

[75] E.L. Lehmann. Some concepts of dependence. *Annals of Mathematical Statistics*, 37(5) :1137–1153, 1966.

[76] F. Liese. Estimates of hellinger integrals of infinitely divisible distributions. *Kybernetika*, 23(3) :227–238, 1987.

[77] M. Mahot. Estimation robuste de la matrice de covariance en traitement du signal. *Thèse de doctorat*, 2012.

[78] R.A. Maronna. Robust m-estimators of multivariate location and scatter. *Annals of Statistics*, 4(1) :51–67, 1976.

[79] R.J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2) :309–323, 1995.

[80] G. Meyer, S. Bonnabel, and R. Sepulchre. Regression on fixed-rank positive semi-definite matrices : A riemannian approach. *Journal of Machine Learning Research*, 12 :593–625, 2011.

[81] R.B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics, 2006.

[82] W. Newey and R. Smith. Higher order properties of gmm and generalized empiracal likelihood estimators. *Econometrica*, 72(1) :219–255, 2004.

[83] E. Ollila, D. Tyler, V. Koivunen, and V. Poor. Complex elliptically symmetric distributions : Survey, new results and applications. *IEEE Transactions on signal processing*, 60(11) :5597–5625, 2012.

[84] Y. Ollivier, L. Arnold, A. Auger, and N. Hansen. Information-geometric optimization algorithms : a unifying picture via invariance principles. *Technical Report*, 2011.

[85] A. Owen. Empirical likelihood ratio confidence regions. *Annals of Statistics*, 18(1) :90–120, 1990.

[86] G. Pailloux. Estimation structurée de la covariance du bruit en détection adaptative. *Thèse de doctorat*, 2010.

[87] E. Parzen. Nonparametric statistical modelling. *Journal of the American Statistical Association*, 74(365) :105–121, 1979.

[88] F. Pascal, P. Forster, J.P. Ovarlez, and P. Larzabal. Performance analysis of covariance matrix estimates in impulsive noise. *IEEE Transatcions on signal processing*, 56(6) :2206–2217, 2008.

[89] M. Pavon and A. Ferrante. On the geometry of maximum entropy problems. *SIAM Review*, 55(3) :415–439, 2013.

[90] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1) :41–66, 2006.

[91] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[92] H. Rohling. Radar cfar thresholding in clutter and multiple target situations. *IEEE Transactions on Aerospace and Electronic Systems*, 19 :608–621, 1983.

[93] M. Rosenblatt. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23(3) :470–472, 1952.

[94] W. Rudin. *Functional Analysis*. Second Edition. McGraw-Hill, 1991.

[95] I.J. Schoenberg. Metric spaces and completely monotone functions. *Annals of Math.*, 39 :811–841, 1938.

[96] T. Sei. Gradient modeling for multivariate quantitative data. *Annals of the Institute of Statistical Mathematics*, 63(4) :675–688, 2011.

[97] R. Serfling. Quantile functions for multivariate analysis : approaches and applications. *Statistica Neerlandica*, 56(2) :214–232, 2002.

[98] R. Serfling. Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardization. *Journal of Nonparametric Statistics*, 22(7) :915–936, 2010.

[99] R. Serfling and P. Xiao. A contribution to multivariate l-moments : L-comoment matrices. *Journal of Multivariate Analysis*, 98(9) :1765–1781, 2007.

[100] S.M. Stigler. Linear functions of order statistics with smooth weight functions. *Annals of Statistics*, 2 :676–693, 1974.

[101] P. Stoica and R.L. Moses. *Spectral Analysis of Signals*. Prentice Hall, 2005.

[102] W. Trench. An algorithm for the inversion of finite toeplitz matrices. *J. Soc. Indust. Appl. Math.*, 12(3) :515–522, 1964.

[103] J.W. Tukey. Which part of the sample contains the information ? *Proceedings of the National Academy of Sciences*, 53 :127–134, 1965.

[104] D. Tyler. A distribution-free m-estimator of multivariate scatter. *Annals of Statistics*, 15(1) :234–251, 1987.

[105] D. Tyler. Statistical analysis for the angular central gaussian distribution on the sphere. *Biometrika*, 74(3) :579–589, 1987.

[106] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[107] S. Verblunsky. On positive harmonic functions : a contribution to the algebra of fourier series. *Proc. London Math. Soc.*, 38(1) :125–157, 1935.

[108] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. Amer. Math. Soc., 2003.

[109] C. Villani. *Optimal transport, old and new*. Grundlehren der Mathematischen Wissenschaften, 338, Springer, 2009.

[110] K.D. Ward, R.J.A. Tough, and S. Watts. *Sea Clutter : Scattering, the K Distribution and Radar Performance*. IET Radar, Sonar and Navigation Series 20, 2006.

[111] L. Yang. Médianes de mesures de probabilité dans les variétés riemanniennes et applications à la détection de cibles radar. *Thèse de doctorat*, 2012.

[112] S.S. Yang. General distribution theory of the concomitants of order statistics. *Annals of Statistics*, 5(5) :996–1002, 1977.

[113] Y. Zuo and R. Serfling. General notions of statistical depth function. *Annals of Statistics*, 28(2) :461–482, 2000.