Thèse de Doctorat

présentée par Zhen Wai Olivier HO

en vue de l'obtention du grade de
Docteur de l'Université Bourgogne Franche-Comté
Spécialité Mathématiques et Applications

Contributions aux algorithmes stochastiques pour le Big Data et à la théorie des valeurs extrêmes multivariées.

Thèse soutenue le 04 octobre 2018, devant le jury composé de :

Yacouba Boubacar-Mainassara	Univ. Franche-Comté	Examinateur
Stéphane Chrétien	NPL	Directeur de thèse
Clément Dombry	Univ. Franche-Comté	Directeur de thèse
Laurent Gardes	Univ. Strasbourg	Rapporteur et Président du jury
Anne Sabourin	Télécom ParisTech	Examinatrice
Joseph Salmon	Univ. Montpellier	Rapporteur

A mes parents, à mon frère, ad infinitum ...

Remerciements

Je souhaite remercier mes directeurs de thèse Clément Dombry et Stéphane Chrétien pour leurs encadrements. C'était un grand plaisir de travailler avec eux. Leurs conseils et supports m'ont permis de mener à bien ces travaux. Je remercie Joseph Salmon et Laurent Gardes pour avoir accepté de rapporter ma thèse. Leurs conseils et suggestions m'ont permis d'améliorer ma thèse. Parallèlement, je souhaite aussi remercier Anne Sabourin et Yacouba Boubacar Mainassara pour avoir accepté de faire partie de mon jury.

Je dois ma gratitude au Laboratoire de Mathématique de Besançon qui a fourni un cadre de travail agréable le long de ma thèse. Je tiens à remercier toute l'équipe de probabilités et statistiques, et en particulier Yacouba pour son support. Merci également à Ulrich Razafison et Rolland Julien Yves pour les pause-cafés où j'ai beaucoup appris.

Mercis à toute l'équipe de doctorants. Je tiens à remercier en particulier Johann Cuenin, Quentin Richard, Ohtman Kadmiri, Aline Moufleh et Tianxiang Gou pour les bons moments qu'on a passé ensemble comme les soirées jeux ou restaurants (avec Julien).

Sans plus d'explications, je remercie mes parents Henri et Pauline ainsi que mon frère Christian. Finalement, je tiens à exprimer ma gratitude à ceux qui d'une manière ou d'une autre m'ont poussé à aller plus loin et ont fait de moi ce que je suis.

Contents

1 Introduction générale **1**
1.1 Généralités sur la théorie des valeurs extrêmes 1
1.1.1 Théorie des valeurs extrêmes univariée 1
1.1.2 Théorie des valeurs extrêmes multivariée 4
1.2 Résultats obtenus dans la partie I 6
1.2.1 Chapitre 2: Simple models for multivariate regular variations 6
1.2.2 Chapitre 3: On the Hüsler-Reiss Pareto distribution 9
1.2.3 Chapitre 4: Numerical study 14
1.3 Généralités sur le machine learning 15
1.3.1 Algorithmes de descente de gradient 18
1.3.2 Généralités sur les algorithmes de gradient stochastique 18
1.3.3 Généralités sur l'acquisition comprimée 24
1.4 Résultats obtenus dans la partie II 29
1.4.1 Chapitre 5: Feature selection in weakly coherent matrices 29
1.4.2 Chapitre 6: Small coherence implies the weak Null Space Property . . 32
1.4.3 Chapitre 7: Incoherent submatrix selection via approximate independence sets in scalar product graphs 33
1.4.4 Chapitre 8: Average performance analysis of the projected gradient method for online PCA 35

I Partie 1 **38**

2 Simple models for multivariate regular variations **39**
2.1 Introduction 39
2.2 A simple model for multivariate regular variations 40
2.2.1 Preliminaries on multivariate regular variations 40
2.2.2 A multivariate version of Breiman Lemma 42
2.2.3 A copula point of view 44
2.2.4 Examples 45
2.2.5 Non standard regular variation 51

3 On the Hüsler-Reiss Pareto distribution **52**
3.1 Introduction . . . 52
3.2 The Hüsler-Reiss Pareto model . . . 52
3.2.1 Definition and transformation properties . . . 52
3.2.2 Exponential family properties . . . 55
3.2.3 Simulation of HR-Pareto random vectors . . . 59
3.2.4 Maximum likelihood inference . . . 60
3.3 The generalised Hüsler-Reiss Pareto model . . . 64
3.3.1 Definition and transformation properties . . . 64
3.3.2 Maximum likelihood inference . . . 66
3.3.3 Optimising the likelihood . . . 70
3.3.4 A likelihood ratio test for $\alpha_1 = \cdots = \alpha_d$ . . . 72

Appendices
3.A Lemmas . . . 74
3.B Argmax theorem . . . 75

4 Numerical study **76**
4.1 Introduction . . . 76
4.2 Numerical simulation: bias and variance in the exact simulation case . . . 76
4.3 Numerical simulation: bias and variance in the domain of attraction simulation case . . . 78

II Partie 2 **81**

5 Feature selection in weakly coherent matrices **82**
5.1 Introduction . . . 82
5.1.1 Background on singular value perturbation . . . 82
5.1.2 Previous approaches to column selection . . . 82
5.1.3 Coherence . . . 83
5.1.4 Contribution of the paper . . . 83
5.2 Main results . . . 84
5.2.1 Appending one vector: perturbation of the smallest non zero eigenvalue 84
5.2.2 Successive perturbations . . . 86
5.3 A greedy algorithm for column selection . . . 86
5.4 Numerical experiments . . . 86
5.4.1 Extracting representative time series . . . 86
5.4.2 Extracting representative images from a dataset . . . 88
5.4.3 Comparison with CUR . . . 89
5.5 Conclusion and perspectives . . . 89

Appendices
5.A Interlacing and the characteristic polynomial . . . 91

5.B Proof of Corollary 5.2 91

6 Small coherence implies the weak Null Space Property 94
6.1 Introduction 94
6.1.1 Motivation 94
6.1.2 Goal of the paper 95
6.1.3 Additional notation 96
6.2 Background 96
6.2.1 Weak NSP and weak RIP 96
6.2.2 On the relationship between RIP and NSP 97
6.2.3 On the relationship between the Coherence and weak-RIP 97
6.2.4 The Gershgorin bound 98
6.3 Main results: small coherence implies weak-NSP 98
6.4 Conclusion 101

Appendices
6.A Technical lemmæ 102
6.A.1 Some perturbation results 102
6.A.2 Appending one vector: perturbation of the smallest non zero eigenvalue 102
6.A.3 Appending one vector: perturbation of the largest eigenvalue 104
6.A.4 Successive perturbations 105
6.A.5 Bounding scalar products 108

7 Incoherent submatrix selection via approximate independence sets in scalar product graphs 109
7.1 Introduction 109
7.2 Incoherent submatrix extraction as an approximate independent set computation 110
7.3 Relaxing on the sphere: a new extraction approach 110
7.3.1 The spectral estimator 110
7.3.2 Theoretical guarantees 111
7.4 Conclusion and future works 113

Appendices
7.A Minimising quadratic functionals on the sphere 114
7.A.1 A semi-explicit solution 114
7.A.2 Bounds on μ 115
7.A.3 ℓ_∞ perturbation of the linear term 116
7.A.4 Neuberger's theorem 118

8 Average performance analysis of the projected gradient method for online PCA 119
8.1 Introduction 119
8.1.1 Background 119
8.1.2 Our contribution 120

8.1.3 Organisation of the paper 120
8.2 Main results 120
8.2.1 Presentation of the problem and prior result 120
8.2.2 The stochastic projected gradient algorithm 121
8.2.3 Main theorem 122
8.3 Proof of the Theorem 8.9 122
8.4 Implementation 124
8.4.1 Choosing the learning rate 124
8.4.2 Numerical experiment 126
8.5 Conclusion 126

Appendices
8.A Technical lemmæ 127

9 Perspectives **131**
9.1 Perspectives suivant les travaux rencontrés dans la partie I 131
9.2 Perspectives suivant les travaux présentés dans la partie II 132

Liste des publications **133**

Bibliographie **134**

Chapter 1

Introduction générale

1.1 Généralités sur la théorie des valeurs extrêmes

Le problème d'estimation de la fréquence d'événements extrêmes est un problème qui a de nombreuses applications. Ainsi pour mettre en contexte ce problème, dans une ère où le climat est plus que jamais affecté par les activités de l'homme, on considère en exemple l'estimation de la fréquence d'événements météorologiques extrêmes comme les sécheresses en Afrique, les ouragans aux États-Unis et les typhons en Asie qui ont des réalités économiques mais aussi humaines.

1.1.1 Théorie des valeurs extrêmes univariée

Historiquement, l'approche de la théorie des valeurs extrême s'orientait vers le comportement du maximum de variables aléatoires indépendantes et identiquement distribuées (i.i.d.). Ainsi, Fisher et Tippett [85] ont établi les premiers résultats fondateurs sur la loi limite des maximums. Gnedenko [91] a ensuite étendu ces résultats avec le théorème de Gnedenko-Fisher-Tippett qui donne les lois limites non-dégénérées possibles pour les maximums.

Theorem 1.1 (Gnedenko-Fisher-Tippett)**.** *Soit $X_1, \ldots, X_n$ variables aléatoires i.i.d. avec fonction de répartition commune F et $M_n = \max(X_1, \ldots, X_n)$. Supposons qu'il existe des suites $(a_n)_n, (b_n)_n$ avec $a_n > 0$ et $b_n \in \mathbb{R}$ telles que*

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{(M_n - b_n)}{a_n} \leq x\right] = \lim_{n\to\infty} (F(a_n x + b_n))^n = G(x) \tag{1.1}$$

où G est non-dégénéré. Alors, à une constante de position et d'échelle près, G est de type de l'une des trois classes suivantes :

- *Fréchet de paramètre $\alpha > 0$ avec une fonction de répartition de la forme*

$$\Phi_\alpha(x) := \exp(-x^{-\alpha}) 1_{x\geq 0};$$

- *Gumbel avec une fonction de répartition de la forme*

$$\Lambda(x) := \exp(-e^{-x});$$

- *Weibull négative de paramètre $\alpha > 0$ avec une fonction de répartition de la forme*

$$\Psi_\alpha(x) := \begin{cases} \exp(-(-x)^\alpha) & , \quad x < 0 \\ 1 & , \quad x \geq 0. \end{cases}$$

Les lois limites ont été paramétrées par Jenkinson [112] en une seule famille, appelée "Generalised Extreme Value" (GEV), donnée par la densité

$$G(z) = \exp\left[-\left\{1 + \gamma\left(\frac{z-\mu}{\sigma}\right)\right\}^{-1/\gamma}\right] \tag{1.2}$$

définie sur $\{z \in \mathbb{R} : 1 + \gamma(z-\mu)/\sigma > 0\}$ avec $-\infty < \mu < \infty, \sigma > 0$ et $-\infty < \eta < \infty$. Le paramètre γ est appelé indice de valeur extrême. La loi limite est de type Fréchet pour $\gamma > 0$, de type Weibull négative pour $\gamma < 0$ et de type Gumbell pour $\gamma = 0$. Lorsque (1.1) a lieu avec la fonction limite

$$G_\gamma(z) = \exp(-(1+\gamma z)^{-1/\gamma})$$

on dit que F est dans le domaine d'attraction de G_γ et on note $F \in MDA(G_\gamma)$.

Le problème de trouver les hypothèses de régularité sur les queues de distribution pour avoir une telle convergence est résolu par les travaux de Gnedenko [91] qui donnent les domaines d'attraction pour les lois GEV avec $\gamma \neq 0$. On citera aussi de Haan [97] pour la caractérisation du domaine d'attraction de la loi Gumbel ainsi que ses reformulations des conditions d'appartenance au domaine d'attraction des lois GEV en terme de variations régulières étendues.

Theorem 1.2 (de Haan)**.** *Pour $\gamma \in \mathbb{R}$, $F \in MDA(G_\gamma)$, si et seulement si*

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma}, \quad x > 0, \tag{1.3}$$

où U est l'inverse continue à gauche de $1/(1-F)$ ($U = (1/(1-F))^{\leftarrow}$) et a est une fonction positive. Lorsque $\gamma = 0$, le terme de droite est interprété comme $\log x$.

On notera tout particulièrement le domaine d'attraction de la loi de Fréchet

$$\Phi_\alpha(z) = \exp(-z^{-\alpha})1_{z>0}, \tag{1.4}$$

que l'on peut formuler simplement comme une condition de variation régulière : $1 - F$ doit varier régulièrement en ∞ avec indice $-\alpha$, c'est-à-dire

$$\lim_{u\to\infty} \frac{1 - F(ux)}{1 - F(u)} = x^{-\alpha}, \quad x > 0.$$

D'un autre côté, plus récemment, la théorie s'intéresse aux comportements des excès au-dessus d'un seuil. Les travaux fondateurs sont dus à Balkema et De Haan[7], Pickands [137]. Le théorème de Pickands-Balkema-de Haan donne alors la loi limite non-dégénérée des excès au-dessus d'un seuil comme étant les lois "Generalised Pareto" (GP). La définition suivante met au clair la notion d'excès.

Definition 1.1. *Soit X une variable aléatoire avec fonction de répartition F et x_F le sup fini ou infini du support de X. Alors, pour $u < x_F$, la fonction*

$$F_u(x) = \mathbb{P}(X - u \leq x | X > u), \quad x \geq 0,$$

est la fonction de répartition des excès de X au dessus de u.

Ainsi défini, le théorème suivant donne la limite en loi des excès au dessus d'un seuil pour des distributions appartenant au domaine d'attraction d'une loi GEV.

Theorem 1.3 (Pickands-Balkema-de Haan)**.** *Soit X une variable aléatoire avec fonction de répartition F. Et soit $\gamma \in \mathbb{R}$ alors $F \in MDA(G_\gamma)$ si et seulement si*

$$\lim_{u \to x_F} \sup_{0 < x < x_F - u} |F_u(x) - H_{\gamma,\beta(u)}(x)| = 0 \tag{1.5}$$

avec β une fonction positive et $H_{\gamma,\beta}$ est la fonction de répartition de la loi Pareto généralisée

$$H_{\gamma,\beta}(x) = 1 - \left(1 + \gamma \frac{x}{\beta}\right)^{-1/\gamma}, \quad 1 + \gamma x/\beta > 0.$$

Par ailleurs, la construction du point de vue des processus ponctuels a été introduite par Dwass [78] et Lamperti [119] sous la notion de processus extrémal. D'un point de vue plus appliqué, l'approche des maximums par bloc (BM) profite de la théorie construite sur la distribution limite des maximums afin de modéliser les événements extrêmes. L'idée étant que la distribution des observations dans le bloc appartient au domaine d'attraction d'une loi GEV de sorte que le maximum du bloc suit approximativement une loi GEV dont on pourra estimer les paramètres. D'un point de vue statistique, plusieurs estimateurs ont été proposés comme l'estimateur du maximum de vraisemblance (MLE) et les estimateurs des moments pondérés par probabilité (PWM) [104]. Sous des conditions du second ordre, de Haan et Ferreira [84] ont obtenu la normalité asymptotique des estimateurs PWM (avec $\gamma < 1/2$). Sous des conditions similaires, Dombry et Ferreira [70] ont obtenu la normalité asymptotique pour les estimateurs MLE (avec $\gamma > -1/2$). Ainsi, dans le cas de l'estimateur MLE, soit $F \in MDA(G_\gamma)$ avec $\gamma > -1/2$, ce qui est équivalent à la convergence des fonctions inverses

$$\lim_{t \to \infty} \frac{V(tx) - V(t)}{a(t)} = G_\gamma^{\leftarrow} = \frac{x^\gamma - 1}{\gamma}, \quad x > 0,$$

avec $V = -(1/\log F)^{\leftarrow}$ et a une fonction positive. Si de plus, on admet une condition sur la vitesse de convergence, c.-à-d. que pour une fonction A satisfaisant $\lim_{t\to\infty} A(t) = 0$, on a

$$\lim_{t \to \infty} \frac{\frac{V(tx)-V(t)}{a(t)} - \frac{x^\gamma - 1}{\gamma}}{A(t)} = \int_1^x s^{\gamma-1} \int_1^s u^{\rho-1} \mathrm{d}u \mathrm{d}s = H_{\gamma,\rho}(x), \quad x > 0, \rho \leq 0 \tag{1.6}$$

alors, on a le théorème suivant sur la normalité asymptotique de l'estimateur MLE:

Theorem 1.4 (Dombry-Ferreira). *Soit $X_1, X_2, \ldots$ i.i.d. avec fonction de répartition commune $F \in MDA(G_\gamma), \gamma > -1/2$ et satisfaisant la condition du second ordre* (1.6). *Soit $k = k_n \to \infty$ le nombre de block et $m = m_n \to \infty$ la taille des blocks de sorte que $\sqrt{k}A(m) \to \lambda \in \mathbb{R}$. Alors il existe une suite d'estimateurs $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} = (\hat{\gamma}_n, \hat{\mu}_n, \hat{\sigma}_n), n \geq 1$, telle que*

$$\lim_{n\to\infty} \mathbb{P}\left[\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} \text{ est un MLE}\right] = 1 \tag{1.7}$$

et

$$\sqrt{k}\left(\hat{\gamma}_n - \gamma, \frac{\hat{\mu}_n - b_m}{a_m}, \frac{\hat{\sigma}_n}{a_m} - 1\right) \xrightarrow{d} \mathcal{N}(\lambda I_{\boldsymbol{\theta}_0}^{-1}\mathbf{b}, I_{\boldsymbol{\theta}_0}^{-1}) \tag{1.8}$$

avec $I_{\boldsymbol{\theta}_0}$ la matrice d'information de Fisher, a_n et b_n les suites normalisantes des maximums partiels et $\mathbf{b} = \mathbf{b}(\gamma, \rho)$ un facteur de biais qui dépend de la condition du second ordre.

Gardes et Girard [89] ont montré que les estimateurs type Pickands pour l'indice de valeur extrême sont asymptotiquement normaux dans le cas $\gamma < -1/2$ et asymptotiquement GEV distribués dans le cas $\gamma > -1/2$.

Une autre approche possible, plus récente, est la modélisation des excès au dessus d'un seuil (PoT). L'idée est simple. Les événements extrêmes sont tellement différents des événements journaliers de sorte que seuls les autres événements extrêmes apportent de l'information. Cette approche repose sur la théorie des excès au dessus d'un seuil. Les estimateurs MLE et PWM ont été proposés et largement étudiés dans la littérature. Ainsi, la normalité asymptotique pour l'estimateur MLE est donnée par Drees et al. [76]. Pour compléter le tableau, la normalité asymptotique dans le cas PWM peut être trouvée dans de Haan et Ferreira [84]. Des comparaisons numériques ont été faites pour contraster les approches BM/PoT et PWM/MLE (Dombry et Ferreira [70], Ferreira et de Haan [84], etc). Un certain consensus se dresse sur le sujet avec la méthode PoT qui semble plus efficace que la méthode BM même si la méthode PoT requiert en moyenne plus d'observations. La combinaison MLE/PoT obtient la meilleure erreur quadratique moyenne optimale asymptotique. On se référera à Beirlant [18] pour une revue plus poussée sur l'approche statistique.

1.1.2 Théorie des valeurs extrêmes multivariée

Les motivations pour une extension multivariée de la théorie des valeurs extrêmes sont diverses et variées. Par exemple, on peut s'intéresser à l'étude spatiale d'événements météorologiques extrêmes. Ou encore, en finance, une question naturelle concerne la dépendance entre les retours extrêmes de produits financiers. Ainsi Tiago de Oliveira [169][170][167], Geffroy [90], Sibuya [156] se sont rapidement intéressés au cas bivarié.

L'extension au cadre multivarié n'est pas une simple transposition de la théorie univariée. Ainsi, de nombreux problèmes sont propres au cadre multivarié. L'obstacle qui apparaît immédiatement revient dans la définition même d'extrême vu qu'il n'y a pas de manière naturelle d'ordonner des observations multivariées (Barnett [15]). Par la suite, on définit

le maximum dans le cas multivarié comme étant le maximum composante par composante, c-à-d que pour $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\mathbf{x} \vee \mathbf{y} := (x_1 \vee y_1, \ldots, x_d \vee y_d). \tag{1.9}$$

La notion de dépendance apparaît naturellement dans le cadre multivarié. L'approche naturelle consiste à traiter les marginales puis, après une normalisation des marginales, à étudier la dépendance. Ainsi, de Haan et Resnick [98] ont obtenu, en supposant sans perte de généralité que les marginales soient Fréchet distribuées, une caractérisation des lois extrêmes multivariées sous le terme de représentation spectrale. Ce résultat utilise le fait que la classe des lois extrêmes multivariées coïncide avec la classe des distributions max-stable multivariées, qui est une sous classe des lois max-infiniment-divisible [8], ce qui donne alors une autre caractérisation en terme de mesure exponentielle. Le théorème est le suivant :

Theorem 1.5 (de Haan-Resnick)**.** *Soit G une loi extrême multivariée à marginales Fréchet unitaire. Alors il existe une mesure μ sur $[0,\infty)^d \setminus \{0\}$ homogène d'ordre -1, c-à-d telle que*

$$\mu(uA) = u^{-1}\mu(A), \quad A \subset [0,\infty)^d \setminus \{0\} \quad \textit{Borélien},$$

de sorte que

$$G(\mathbf{x}) = \exp\left(-\overline{\mu}(x)\right), \quad \mathbf{x} \in (0,\infty)^d, \tag{1.10}$$

avec $\overline{\mu}$ la fonction de survie de μ définie par

$$\overline{\mu}(\mathbf{x}) = \mu([\mathbf{0}, \mathbf{x}]^c) < \infty, \quad \mathbf{x} \in (0,\infty)^d.$$

Une autre caractérisation est donnée par Huang [105] qui introduit le terme de fonction de dépendance de queue stable (*stable tail dependence function*).

Finalement, une autre représentation populaire est celle des copules qui ont été introduites pour décrire la structure de dépendance de lois multivariés par Sklar [158]. Ce choix correspond au cas où les marginales sont uniformément distribuées.

Definition 1.2. *Une copule C est la fonction de répartition d'un vecteur multivarié $\mathbf{Z} \in [0,1]^d$ à marginales de loi uniformes sur $[0,1]$.*

Theorem 1.6 (Sklar)**.** *Toute fonction de répartition F sur $\mathbb{R}^d$ avec marginales $F_1, \ldots, F_d$ peut être décomposée en*

$$F(\mathbf{x}) = C(F_1(x_1), \ldots, F_d(x_d)), \quad \mathbf{x} \in \mathbb{R}^d \tag{1.11}$$

où C est une copule. Si F est continue alors C est unique. La copule C_F associée à F est donnée par

$$C_F(\mathbf{u}) = F\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right), \quad \mathbf{u} \in (0,1)^d \tag{1.12}$$

Deheuvels [63][64] a donné la caractérisation des domaines d'attraction sous le point de vue copule.

Theorem 1.7 (Deheuvels)**.** *Une loi multivariée F avec marginales $F_1, \ldots, F_d$ appartient au domaine d'attraction de la loi GEV multivariée G avec marginales $G_1, \ldots, G_d$ si et seulement si*

- $F_i \in MDA(G_i), i = 1, \ldots, d$ *(cf Théorème 1.1).*
- *la copule associée à F est dans le domaine d'attraction de la copule associée à G dans le sens*

$$\lim_{n\to\infty} C_F^n(u_1^{1/n}, \ldots, u_d^{1/n}) = C_G(u_1, \ldots, u_d), \quad \mathbf{u} \in (0,1)^d. \tag{1.13}$$

Marshall et Olkin [126] présentent des analogues au Théorème de Gnedenko dans le cas multivarié sur la caractérisation des domaines d'attraction.

Plus récemment, Coles et Tawn [58], Rootzén et Tajvidi [148] ont réintroduit l'approche des excès au dessus d'un seuil.

Du point de vue de la modélisation, de nombreux modèles paramétriques ont été présentés par Gumbel [95], Hüsler et Reiss [109], Coles et Tawn [58], Brown et Resnick [30], etc. Toute une littérature a été écrite dans ce sens, mais on citera en particulier Tawn [164][159][165][166] pour ses travaux.

Dombry, Engelke et Oesting donnent des algorithmes pour la simulation exacte de processus max-stable multivariés [69] et donnent des conditions sur l'existence d'un estimateur du maximum de vraisemblance local asymptotiquement normal et efficace [67]. Par ailleurs, les récents travaux de Rootzén, Wadsworth et Segers [147][146] se concentrent sur l'aspect statistique et modélisation des lois Pareto généralisées multivariées.

1.2 Résultats obtenus dans la partie I

La première partie regroupe les travaux effectués sous la direction de Clément Dombry. Ces chapitres sont issus d'un article soumis pour publication à Journal of Multivariate Analysis [103] et une première révision est en cours.

Notation vectorielle pour la première partie: on note $\|\cdot\|_\infty$ norme max sur $\mathbb{R}^d$ et $\|\cdot\|$ une norme arbitraire, $\mathbf{1_d} = (1, \ldots, 1)$ est le vecteur avec toute les composantes égales à 1. Les opérations sur les vecteurs sont, sauf mention du contraire, prises composantes par composantes. Le maximum composante par composante de vecteur est noté $\max(\mathbf{x_1}, \mathbf{x_2}) = \mathbf{x_1} \vee \mathbf{x_2}$, la comparaison entre les vecteurs $\mathbf{x_1} \leq \mathbf{x_2}$ est à prendre composante par composante de sorte que $\mathbf{x_1} \not\leq \mathbf{x_2}$ signifie que certaines composantes de $\mathbf{x_1}$ sont plus grandes que les composantes associées de $\mathbf{x_2}$. Pour $\mathbf{x} \in [0,\infty)^d$, on note $[\mathbf{0}, \mathbf{x}]$ le cube $[0, x_1] \times \cdots \times [0, x_d]$ et $[\mathbf{0}, \mathbf{x}]^c = [0,\infty)^d \setminus [\mathbf{0}, \mathbf{x}]$.

1.2.1 Chapitre 2: Simple models for multivariate regular variations

Dans ce chapitre, on donne une construction de vecteurs aléatoires à variations régulières qui nous permet de retrouver les modèles classiques max-stable multivariés rencontrés dans la

littérature.

On rappelle la notion de fonction à variations régulières en $+\infty$ qui sert de base pour construire la notion de variable aléatoire variant régulièrement.

Definition 1.3. *Une fonction mesurable $f : \mathbb{R}^+ \to \mathbb{R}^+$ est dite à variation régulière en ∞ avec indice α et notée $f \in RV_\alpha$ si*

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^\alpha, \quad x > 0. \tag{1.14}$$

On dira alors qu'une variable aléatoire positive X varie régulièrement si sa queue de distribution $1 - F$ est à variation régulière, c'est-à-dire

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^\alpha \quad , \alpha \in \mathbb{R}. \tag{1.15}$$

Étant donné l'espace $M_0(\mathbb{R}^d)$ des mesures boréliennes μ sur $\mathbb{R}^d \setminus \{0\}$ tel que $\mu(\mathbb{R}^d \setminus O)$ est finie pour tout voisinage ouvert O de 0, une suite $\mu_n \in M_0(\mathbb{R}^d)$ converge vers $\mu \in M_0(\mathbb{R}^d)$ si $\int f \mathrm{d}\mu_n \longrightarrow \int f \mathrm{d}\mu$ pour toute fonction f continue, bornée et s'annulant dans un voisinage de 0. On définit alors la notion de variation régulière multivariée d'un vecteur aléatoire $\mathbf{X}$ comme étant la convergence

$$n\mathbb{P}(\mathbf{X}/a_n \in \cdot) \xrightarrow{M_0} \Lambda, \quad n \to \infty \tag{1.16}$$

pour une suite $a_n \to +\infty$ et mesure limite non-dégénérée $\Lambda \in M_0(\mathbb{R}^d)$. Une telle mesure limite Λ a la propriété d'être homogène, c'est-à-dire qu'il existe un réel $\alpha > 0$ tel que

$$\Lambda(uA) = u^{-\alpha}\Lambda(A) \quad u > 0, A \subset \mathbb{R}^d \setminus \{0\} \text{ Borélien.} \tag{1.17}$$

Dans le cas de vecteur aléatoires à composantes positives $\mathbf{X}$, la notion de variation régulière sur $[0, \infty)^d$ est caractérisée par la variation régulière de la fonction de survie multivariée. C'est-à-dire que pour F la fonction de répartition de $\mathbf{X}$, on a

$$\lim_{u\to+\infty} \frac{1 - F(u\mathbf{x})}{1 - F(u\mathbf{1_d})} = V(\mathbf{x}), \quad \mathbf{x} \in [0, \infty)^d \setminus \{\mathbf{0}\}, \tag{1.18}$$

où la fonction limite V est donnée par la mesure du complémentaire du pavé $[0, x]^d$ qu'on notera $\Lambda([\mathbf{0}, \mathbf{x}]^c)$ et $\mathbf{1_d} = (1, \ldots, 1) \in \mathbb{R}^d$.

Pour construire des vecteurs aléatoires variant régulièrement sur $\mathbb{R}^d \setminus \{\mathbf{0}\}$, une possibilité est de considérer le produit $X = R\mathbf{Z}$ entre une variable aléatoire R positive et à variation régulière d'indice $\alpha > 0$ et un vecteur aléatoire $\mathbf{Z}$ suffisamment intégrable, par exemple $\alpha + \varepsilon$ intégrable avec $\varepsilon > 0$. Cette construction est donnée par la proposition suivante

Proposition 1.1. *Soit R une variable aléatoire positive et $\mathbf{Z}$ un vecteur d-dimensionnel indépendant de R. Alors, si l'une des deux hypothèses suivantes est vérifiée*

- *la queue de distribution* $1-F$ *de* R *varie régulièrement en* $+\infty$ *avec indice* $-\alpha<0$ *et* $\mathbb{E}[\|\mathbf{Z}\|^{\alpha+\varepsilon}]<\infty$ *pour* $\varepsilon>0$*;*
- $1-F(x)\sim Cx^{-\alpha}$ *lorsque* $x\to\infty$ *avec* $C>0$ *et* $\mathbb{E}[\|\mathbf{Z}\|^{\alpha}]<\infty$*,*

le produit $X=R\mathbf{Z}$ *définit un vecteur aléatoire variant régulièrement sur* $[-\infty,\infty]^d\setminus\{\mathbf{0}\}$ *avec indice* α*. C'est-à-dire*

$$n\mathbb{P}(a_n^{-1}X\in\cdot)\xrightarrow{M_0}\Lambda(\cdot)\quad \textit{dans } M_0(\mathbb{R}^d) \textit{ lorsque } n\to\infty, \tag{1.19}$$

où a_n *est le quantile d'ordre* $1-1/n$ *de* R *et la mesure limite* Λ *est donnée par*

$$\Lambda(A)=\int_0^\infty \mathbb{P}(uZ\in A)\alpha u^{-\alpha-1}\mathrm{d}u,\quad A\subset\mathbb{R}^d\setminus\{\mathbf{0}\} \textit{ Borélien.} \tag{1.20}$$

Par ailleurs, si $\mathbf{Z}$ *est positif, alors le support de* Λ *est donné par* $[0,\infty)^d\setminus\{\mathbf{0}\}$ *et la fonction limite* V *est caractérisée par*

$$V(x):=\Lambda([\mathbf{0},\mathbf{x}]^c)=\mathbb{E}\left[\bigvee_{i=1}^d\left(\frac{Z_i}{x_i}\right)^\alpha\right],\quad \mathbf{x}\in[0,+\infty)\setminus\{\mathbf{0}\}. \tag{1.21}$$

Cette construction peut être vue dans le cadre plus général de la théorie des valeurs extrêmes multivariées comme le produit entre une composante radiale et une composante angulaire, R étant alors la composante radiale. La preuve de la proposition illustre bien ce point de vue, l'idée étant que sur les ensembles de la forme

$$A=\left\{\mathbf{z}\in\mathbb{R}^d:\|\mathbf{z}\|>x,\mathbf{z}/\|\mathbf{z}\|\in B\right\},\quad B\subset\mathcal{S}^{d-1}\text{ Borélien }, x>0$$

qui forment une classe déterminant la convergence, on peut appliquer le lemme de Breiman univarié 2.1 pour obtenir la convergence M_0 et la caractérisation de la mesure limite Λ.

D'un point de vue copule, on trouve une interprétation de la proposition précédente. On considère le modèle de convolution avec un seul facteur commun [117]

$$\mathbf{X}=\alpha E\mathbf{1_d}+\mathbf{Y} \tag{1.22}$$

où $\alpha>0$, E suivant une loi exponentielle et $\mathbf{Y}$ un vecteur $d-$dimensionnel avec $\mathbb{E}[e^{\alpha Y_i}]<\infty, i=1,\dots,d$. On a alors:

Proposition 1.2. *Soit* $C_{\mathbf{X}}$ *la copule associée au vecteur aléatoire* $\mathbf{X}$ *défini par l'équation* (1.22)*. Alors*

$$C_{\mathbf{X}}^n(u_1^{1/n},\dots,u_d^{1/n})\to C_V(u_1,\dots,u_d),\quad (u_1,\dots,u_d)\in[0,1]^d, \tag{1.23}$$

où

$$C_V(u_1,\dots,u_n)=\exp(-V(\sigma_1(-\log u_1)^{1/\alpha},\dots,\sigma_d(-\log u_d)^{1/\alpha}))$$

et

$$\sigma_i^\alpha=\mathbb{E}[e^{\alpha Y_i}]\quad \textit{et}\quad V(\mathbf{x})=\mathbb{E}\left[\bigvee_{i=1}^d\frac{e^{\alpha Y_i}}{x_i^\alpha}\right].$$

L'idée de la preuve est de remarquer qu'en prenant l'exponentielle de $\mathbf{X}$, on retrouve le produit de $\exp(\alpha\mathbf{E})$ et $\exp Y$. Par hypothèse sur $\mathbf{Y}$ et par le fait que $\exp(\alpha\mathbf{E})$ suit une loi α-Pareto alors le produit forme un vecteur positif variant régulièrement. Puis finalement, il faut remarquer que comme l'exponentielle agit composante par composante, la copule de $\exp(\mathbf{X})$ est $C_{\mathbf{X}}$ et $C_{\mathbf{X}}(u_1^{1/n}, \dots, u_d^{1/n})$ est la copule du maximum normalisé de n copies de $\mathbf{X}$ indépendantes. Finalement, C_V est la copule du vecteur α-Fréchet limite.

Une question naturelle se pose sur la caractérisation de la mesure limite lorsque la composante angulaire $\mathbf{Z}$ est à densité $f_{\mathbf{Z}}$. D'où la proposition suivante

Proposition 1.3. *Si $\mathbf{Z}$ a une densité $f_{\mathbf{Z}}$, alors la mesure limite Λ a aussi une densité λ donnée par*

$$\lambda(\mathbf{z}) = \int_0^\infty f_{\mathbf{z}}(\mathbf{z}/u)\alpha u^{-d-\alpha-1}\mathrm{d}u. \tag{1.24}$$

Il devient alors naturel d'étudier la forme de la densité pour des lois classiques multivariées à densité. On retrouve alors des modèles connus comme le modèle max-stable t-extrémal lorsque $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ avec

$$\lambda(\mathbf{z}) = \frac{\alpha}{(2\pi)^{d/2}|\Sigma|^{1/2}}\Gamma\left(\frac{\alpha+d}{2}\right)\left(\frac{\mathbf{z}^\top\Sigma^{-1}\mathbf{z}}{2}\right)^{-(\alpha+d)/2}, \quad \mathbf{z} \in \mathbb{R}^d \setminus \{0\}, \tag{1.25}$$

le modèle max-stable Hüsler-Reiss lorsque $\ln\mathbf{Z} \sim \mathcal{N}(\mathbf{m}, \Sigma)$ avec

$$\lambda(\mathbf{z}) = C\exp\left\{-\frac{1}{2}\log\mathbf{z}^\top Q\log\mathbf{z} + \boldsymbol{l}\log\mathbf{z}\right\}\prod_{i=1}^d z_i^{-1}, \quad \mathbf{z} \in (0,\infty)^d \tag{1.26}$$

où

$$C = \frac{\alpha}{(2\pi)^{(d-1)/2}|\Sigma|^{1/2}\sqrt{\mathbf{1_d}^\top\Sigma^{-1}\mathbf{1_d}}}\exp\left\{-\frac{1}{2}\mathbf{m}^\top\Sigma^{-1}\mathbf{m} + \frac{1}{2}\frac{(\mathbf{m}^\top\Sigma^{-1}\mathbf{1_d} - \alpha)^2}{\mathbf{1_d}^\top\Sigma^{-1}\mathbf{1_d}}\right\},$$

$$Q = \Sigma^{-1} - \frac{\Sigma^{-1}\mathbf{1_d}\mathbf{1_d}^\top\Sigma^{-1}}{\mathbf{1_d}^\top\Sigma^{-1}\mathbf{1_d}}, \tag{1.27}$$

$$\boldsymbol{l} = \left(\mathbf{m}^\top - \frac{\alpha + \mathbf{m}^\top\Sigma^{-1}\mathbf{1_d}}{\mathbf{1_d}^\top\Sigma^{-1}\mathbf{1_d}}\mathbf{1_d}^\top\right)\Sigma^{-1}. \tag{1.28}$$

ainsi que d'autres modèles max-stable.

1.2.2 Chapitre 3: On the Hüsler-Reiss Pareto distribution

Sous le cadre donné précédemment, on retrouve le modèle de Pareto associé au modèle Hüsler-Reiss comme la limite en loi des excès au dessus d'un seuil $\mathbf{a} \in \mathbb{R}^d$

$$\lim_{u\to\infty}\mathbb{P}[u^{-1}\mathbf{X} \not\leq \mathbf{x} | X \not\leq u\mathbf{a}] = \frac{V(\mathbf{x}\vee\mathbf{a})}{V(\mathbf{a})}, \quad \mathbf{x} \in [0,\infty)^d \setminus [\mathbf{0}, \mathbf{a}] \tag{1.29}$$

avec $\mathbf{X}$ dans le domaine d'attraction d'un modèle max-stable Hüsler-Reiss. Partant de la caractérisation de la densité de la loi limite donnée précédemment, on retrouve la densité associée qui définit ainsi le modèle de Hüsler-Reiss-Pareto.

Definition 1.4. *Soient $d \geq 2$, $Q \in \mathbb{R}^{d\times d}$ une matrice symétrique semi-définie positive telle que $\mathrm{Ker}(Q) = \mathrm{vect}(\mathbf{1_d})$, $\boldsymbol{l} \in \mathbb{R}^d$ vérifiant $\boldsymbol{l}^\top \mathbf{1_d} < 0$, et $\mathbf{a} = (a_1, \ldots, a_d) \in (0,\infty)^d$ le seuil. Le modèle Hüsler-Reiss-Pareto sur $[0,\infty)^d \setminus [\mathbf{0}, \mathbf{a}]$ paramétré par $(Q, \boldsymbol{l})$ est défini par la densité*

$$f_{\mathbf{a}}(\mathbf{z}; Q, \boldsymbol{l}) = \frac{1}{C_{\mathbf{a}}(Q,\boldsymbol{l})} \exp\left(-\frac{1}{2}\log \mathbf{z}^\top Q \log \mathbf{z} + \boldsymbol{l}^\top \log \mathbf{z}\right)\left(\prod_{i=1}^d z_i^{-1}\right) 1_{\{\mathbf{z}\not\leq\mathbf{a}\}}, \quad \mathbf{z} \in (0,\infty)^d \tag{1.30}$$

avec $C_{\mathbf{a}}(Q,\boldsymbol{l})$ la constante de normalisation. On note alors $Z \rightsquigarrow \mathrm{HRPar}_{\mathbf{a}}(Q,\boldsymbol{l})$ si le vecteur aléatoire $\mathbf{Z}$ a pour densité $f_{\mathbf{a}}$.

Le cadre donné précédemment permet de relier les paramètres de modèle Hüsler-Reiss à l'indice de variation régulière et aux paramètres de la loi log-normale qui composent le vecteur à variation régulière, par exemple $\boldsymbol{l}^T \mathbf{1}_d$ est égal à $-\alpha$. De la même façon que la loi log-normale, la loi Hüsler-Reiss Pareto est invariant par changement d'échelle, c'est-à-dire

Proposition 1.4. *Soit $\mathbf{Z} \rightsquigarrow \mathrm{HRPar}_a(Q,l)$. Alors*

- *pour tout $\mathbf{u} \in (0,\infty)^d$, $\mathbf{uZ} \rightsquigarrow \mathrm{HRPar}_{\mathbf{ua}}(Q, \boldsymbol{l} + \log \mathbf{u})$, et*
- *pour tout $\beta > 0$, $\mathbf{Z}^\beta \rightsquigarrow \mathrm{HRPar}_{\mathbf{a}^\beta}(\beta^{-2}Q, \beta^{-1}\boldsymbol{l})$.*

Ainsi sous réserve de reparamétrisation, il est toujours possible de se ramener au cas $\mathbf{a} = \mathbf{1_d}$. Par la suite, on considère donc $\mathbf{a} = \mathbf{1_d}$. Le bon cadre qui permet l'étude du modèle de Hüsler-Reiss Pareto est le cadre des familles exponentielles. Ainsi, le résultat principal place le modèle Hüsler-Reiss Pareto dans ce cadre avec le théorème suivant

Theorem 1.8. *Soit E l'espace euclidien $d(d+1)/2$-dimensionnel défini par*

$$E = \left\{(A, \mathbf{b}) \in \mathbb{R}^{d\times d} \times \mathbb{R}^d \ : \ A^\top = A, \ A\mathbf{1_d} = \mathbf{0}\right\}$$

muni du produit scalaire

$$\langle (A, \mathbf{a}), (A', \mathbf{a}')\rangle = \sum_{1\leq i,j\leq d} A_{i,j}A'_{i,j} + \sum_{1\leq k\leq d} a_k a'_k.$$

Soit Θ le sous-ensemble de E défini par

$$\Theta = \left\{(Q, \boldsymbol{l}) \in E \ : \ Q \text{ semi définie positive}, \mathrm{Ker}(Q) = \mathrm{vect}(\mathbf{1_d}), \boldsymbol{l}^\top \mathbf{1_d} < 0\right\}.$$

Pour tout $\mathbf{a} \in (0,\infty)^d$, les lois Hüsler-Reiss Pareto $f_{\mathbf{a}}(\mathbf{z};\theta)_{\theta\in\Theta}$ forment une famille exponentielle complète canonique paramétrée par $\theta = (Q, \boldsymbol{l}) \in \Theta$ et ayant comme statistique suffisante

$$T(\mathbf{z}) = \left(-\frac{1}{2}(\log \mathbf{z} - \overline{\log \mathbf{z}})(\log \mathbf{z} - \overline{\log \mathbf{z}})^\top, \log \mathbf{z}\right),$$

où $\overline{\log \mathbf{z}} = d^{-1}(\mathbf{1_d}^\top \log \mathbf{z})\mathbf{1_d}$.

L'idée de la preuve est de remarquer que comme $\mathbf{1_d}$ appartient au noyau de Q, un changement de variable accompagné du théorème de Fubini nous permet de séparer l'intégrale en deux parties où chacune des deux parties donne les conditions recherchées. Sous le cadre de la théorie des familles exponentielles [14], le calcul du terme de normalisation est important car il nous permet de calculer les moments de la statistique naturelle. Les calculs du terme de normalisation nous permettent aussi d'obtenir une méthode de simulation exacte. Puis, on s'intéresse à l'inférence par l'estimateur du maximum de vraisemblance. Notre principal théorème est le suivant concernant l'existence, l'unicité et la normalité asymptotique du maximum de vraisemblance

Theorem 1.9. *Soient* $\mathbf{a} \in (0,\infty)^d$ *et* $n \geq 1$.

(i) (existence et unicité) Pour des observations $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(n)} \in [\mathbf{0}, \mathbf{a}]^c$*, la log-vraisemblance* $(Q, \boldsymbol{l}) \mapsto L_n(Q, \boldsymbol{l}; \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(n)})$ *est strictement concave sur* Θ*. Un estimateur du maximum de vraisemblance existe si et seulement si*

$$V_n = \frac{1}{n}\sum_{i=1}^{n} \log \mathbf{z}^{(i)} \log \mathbf{z}^{(i)T} - \left(\frac{1}{n}\sum_{i=1}^{n} \log \mathbf{z}^{(i)}\right)\left(\frac{1}{n}\sum_{i=1}^{n} \log \mathbf{z}^{(i)}\right)^{\top}$$

est conditionnellement définie positive dans le sens où $\mathbf{v}^\top V_n \mathbf{v} > 0$ *pour tout* $\mathbf{v} \in \mathbb{R}^d \backslash \{\mathbf{0}\}$ *tel que* $\mathbf{v}^\top \mathbf{1_d} = 0$*. S'il existe, le maximum de vraisemblance* $\hat{\theta}_n^{mle}$ *est l'unique solution de l'équation du score*

$$\frac{\partial \log C_{\mathbf{a}}}{\partial \theta}(\theta) = \overline{T}_n, \quad \theta \in \Theta. \tag{1.31}$$

(ii) (normalité asymptotique) Soit $\theta = (Q, \boldsymbol{l}) \in \Theta$ *et supposons que* $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(n)}$ *soient générés suivant la loi* $\mathrm{HRPar}_a(Q, \boldsymbol{l})$*. Alors, pour* $n \geq d-1$*, il existe presque sûrement un unique estimateur du maximum de vraisemblance* $\hat{\theta}_n^{mle}$ *qui est asymptotiquement normal et efficace, c'est-à-dire*

$$\sqrt{n}(\hat{\theta}_n^{mle} - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1}), \quad \textit{lorsque } n \to \infty,$$

où $I(\theta)$ *est la matrice d'information de Fisher*

$$I(\theta) = -\frac{\partial^2 \log C_{\mathbf{a}}}{\partial \theta \partial \theta^\top}(\theta).$$

L'idée de la preuve repose sur la théorie générale des familles exponentielles [14] qui donne une caractérisation de l'existence et l'unicité du maximum de vraisemblance par l'appartenance de la statistique suffisante $\overline{T}_n$ à l'intérieur de la fermeture convexe du support de la statistique T. Ainsi, on détermine $\mathrm{int}(\overline{\mathrm{conv(S)}})$ et on montre $\overline{T}_n \in \mathrm{int}(\overline{\mathrm{conv(S)}})$ si et seulement si V_n est conditionnellement définie positive. La seconde partie du théorème est un résultat général pour les familles exponentielles complètes.

Finalement, on s'intéresse à l'extension du modèle Hüsler-Pareto aux variations régulières non-standard. Cette notion qui avait été introduite par Resnick [141] correspond au cas où les marginales ont des indices de queues différents. On définit alors le modèle Hüsler-Reiss Pareto par

Definition 1.5. *Soit $d \geq 2$ et Θ l'ensemble défini par*

$$\Theta = \big\{(\boldsymbol{\alpha}, Q, \boldsymbol{l}) \in (0,\infty)^d \times \mathbb{R}^{d\times d} \times \mathbb{R}^d : Q \text{ symétrique semi-définie positive }, \\ \mathrm{Ker} Q = \mathrm{vect}(\mathbf{1_d}) \text{ et } \boldsymbol{l}^\top \mathbf{1_d} = -1\big\}$$

Alors pour le seuil $\mathbf{a} \in (0,\infty)^d$, le modèle Hüsler-Reiss Pareto généralisé sur $[0,\infty)^d \setminus [\mathbf{0}, \mathbf{a}]$ paramétré par $\theta = (\boldsymbol{\alpha}, Q, \boldsymbol{l})$ est définie par la densité

$$f_{\mathbf{a}}(\mathbf{z};\theta) = \frac{1}{C_{\mathbf{a}}(\theta)} \exp\left(-\frac{1}{2}\log \mathbf{z}^\top D_{\boldsymbol{\alpha}} Q D_{\boldsymbol{\alpha}} \log \mathbf{z} + \boldsymbol{l}^\top D_{\boldsymbol{\alpha}} \log \mathbf{z}\right) \left(\prod_{i=1}^d z_i^{-1}\right) 1_{\{\mathbf{z} \not\leq \mathbf{a}\}}$$

où $C_{\mathbf{a}}(\theta)$ est la constante de normalisation et $D_{\boldsymbol{\alpha}}$ la matrice diagonale ayant pour diagonale α.

On remarquera la condition supplémentaire $\boldsymbol{l}^\top \mathbf{1_d} = -1$ que l'on pose pour identifier le modèle. En effet pour $\lambda > 0$, la densité est invariante par rapport au changement de variable $(\boldsymbol{\alpha}, Q, \boldsymbol{l}) \mapsto (\lambda\boldsymbol{\alpha}, \lambda^{-1/2}Q, \lambda^{-1}\boldsymbol{l})$. Dans le cas où tous les indices de variations régulières α_i sont égaux alors on retrouve le modèle Hüsler-Reiss Pareto. Par ailleurs, comme le modèle Hüsler-Reiss Pareto, le modèle dit généralisé est aussi stable par changement d'échelle. Cette propriété sera revisitée plus loin lorsqu'on abordera les procédures d'optimisation dans le cadre d'inférence par le maximum de vraisemblance. Les arguments qui ont permis l'étude de l'estimateur de vraisemblance dans le cas non généralisé ne peuvent pas être utilisés dans le cas généralisé. En effet, la famille des distributions Hüsler-Reiss Pareto généralisée forme une famille exponentielle courbée avec statistique minimale suffisante T donnée par

$$T(\mathbf{z}) = (\log \mathbf{z} \log \mathbf{z}^\top, \log \mathbf{z})$$

et l'ensemble Θ des paramètres n'est pas strictement inclus dans l'intérieur de l'espace naturel des paramètres associé à cette famille. Néanmoins, en montrant la différentiabilité en moyenne quadratique du modèle statistique $\{f_{\mathbf{1_d}}(\theta;\mathbf{z}), \theta \in \Theta\}$ et en utilisant une expansion uniforme du processus de vraisemblance au voisinage du paramètre θ_0 combinée avec le théorème Argmax ([176], Corollaire 5.58, voir Appendice 3.B), nous parvenons à étudier les propriétés asymptotiques de l'estimateur du maximum de vraisemblance. D'où le résultat suivant

Theorem 1.10. *Soient $\theta_0 \in \Theta$ avec I_{θ_0} définie positive et $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots$ i.i.d suivant une loi* $\mathrm{HRPar}_{\mathbf{a}}(\theta_0)$. *Alors, il existe un estimateur du maximum de vraisemblance $\hat{\theta}_n^{mle}$ qui est asymptotiquement normal et efficace, c'est-à-dire*

$$\sqrt{n}(\hat{\theta}_n^{mle} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_{\theta_0}^{-1}) \quad \text{lorsque } n \to \infty.$$

L'idée de la preuve est que sous réserve que I_{θ_0} soit définie positive, le développement de Taylor d'ordre 2 sur voisinage compact de θ_0 implique la concavité stricte du processus de vraisemblance local avec forte probabilité. Puis, en montrant que la suite des maximums $\hat{h}_n$ du processus de vraisemblance est tendue, on applique le théorème Argmax [176] qui

nous donne le résultat. Néanmoins, on a montré que la log-vraisemblance est strictement concave sur un voisinage de θ_0 et non pas globalement. On remarquera par contre que la log-vraisemblance est biconcave, c'est-à-dire que les fonctions partielles $\boldsymbol{\alpha} \mapsto L_n(\boldsymbol{\alpha}, Q, \boldsymbol{l})$ et $(Q, \boldsymbol{l}) \mapsto L_n(\boldsymbol{\alpha}, Q, \boldsymbol{l})$ sont concaves. On propose alors un estimateur des moments pour initialiser une routine d'optimisation de la log-vraisemblance. En utilisant la loi forte des grands nombres, on montre la consistance forte de cet estimateur puis le théorème central limite combiné avec la delta méthode implique la normalité asymptotique que l'on résume dans la proposition suivante

Theorem 1.11. *Soient $\theta = (\boldsymbol{\alpha}, Q, \boldsymbol{l}) \in \Theta$ et $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots$ i.i.d suivant une loi $\mathrm{HRPar}_{\mathbf{a}}(\theta)$. Pour $j = 1, \ldots, d$, on définit*

$$N_{n,j} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Z_j^{(i)} > 1} \quad \textit{et} \quad O_{n,j} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Z_j^{(i)} > 1} \log Z_j^{(i)}.$$

Alors l'estimateur $\hat{\theta}_0 = (\hat{\boldsymbol{\alpha}}_{\mathbf{0}}, \hat{Q}_0, \hat{\boldsymbol{l}}_{\boldsymbol{n}})$ défini par

$$\hat{\boldsymbol{\alpha}}_0 = (N_{n,j}/O_{n,j})_{1 \leq j \leq d} \quad \textit{et} \quad (\hat{Q}_0, \hat{\boldsymbol{l}}_{\mathbf{0}}) = \mathrm{argmax}_{Q,\boldsymbol{l}} L_n(\hat{\boldsymbol{\alpha}}_{\mathbf{0}}, Q, \boldsymbol{l})$$

est fortement consistant et asymptotiquement normal.

Finalement, on montre que la suite d'estimateurs obtenus par une routine de maximisation alternée initialisée par l'estimateur des moments converge presque sûrement vers l'unique maximiseur de la log-vraisemblance dans le voisinage du vrai paramètre. En effet, le théorème de Prohorov implique que l'estimateur des moments $\hat{\theta}_0$ appartient avec forte probabilité à un voisinage de θ. En utilisant la propriété de biconcavité de la log-vraisemblance, on obtient que chaque itéré de l'algorithme de maximisation alternée reste dans le voisinage de θ. Pour terminer, on propose un test du rapport de vraisemblance pour l'hypothèse $H_0 : \alpha_1 = \cdots = \alpha_d$. Encore une fois, on utilise un développement du processus de vraisemblance local pour obtenir le résultat suivant :

Theorem 1.12. *Soit $\theta_0 = (\boldsymbol{\alpha}, Q, \boldsymbol{l}) \in \Theta$ avec $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$. Soit $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(n)}$ i.i.d. de loi $\mathrm{HRPar}(\theta_0)$. On note $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance pour le modèle Hüsler-Reiss Pareto généralisé et $\hat{\theta}_0$ l'estimateur du maximum de vraisemblance dans le modèle Hüsler-Reiss Pareto. On définit alors la différence des log-vraisemblance par*

$$\Delta_n = L_n(\hat{\theta}_n) - L_n(\hat{\theta}_0).$$

Alors, sous l'hypothèse nulle $\alpha_1 = \cdots = \alpha_d$, $2\Delta_n$ converge en loi vers une loi du khi-deux à $d-1$ degrés de liberté, c'est-à-dire

$$2(L_n(\hat{\theta}_n) - L_n(\hat{\theta}_0)) \xrightarrow{d} \chi^2(d-1).$$

1.2.3 Chapitre 4: Numerical study

Dans ce chapitre, on illustre les résultats donnés dans le chapitre précédent par des études de simulations. On étudie ainsi les propriétés de l'estimateur du maximum de vraisemblance dans différents cadres de simulations. L'un des cadres proposés est le cas de la simulation exacte. Plusieurs études sont possibles comme l'étude de l'effet de la dimension d sur les estimateurs. Comme le nombre de paramètre du modèle est égal à $d(d+1)/2$, on choisit de comparer les estimateurs pour $\alpha = -\sum l_i$ et $Q_{11} = 1 - \sum_{i>1} Q_{1_i}$ et on impose une structure "symétrique" aux paramètres. Ainsi on fixe les paramètres $Q = I_d - \mathbf{1_d}\mathbf{1_d}^\top/d$ et $\boldsymbol{l} = -\alpha/d\mathbf{1_d}$. Dans ce cas, puisque les l_i sont fixés et dépendent uniquement de α, on étudie aussi l'effet de α sur l'estimation. Puis, comme les résultats sur l'estimateur du maximum de vraisemblance sont asymptotiques, on fait varier la taille de l'échantillon n pour observer le comportement de l'estimateur sur des échantillons finis. Finalement, on répète l'expérience 1000 fois pour obtenir un échantillon Monte-Carlo qui nous donne les résultats suivants

		$\alpha = 0.5$				$\alpha = 1.0$				$\alpha = 1.2$			
		$\hat{\alpha}$		$\hat{Q}_{11}$		$\hat{\alpha}$		$\hat{Q}_{11}$		$\hat{\alpha}$		$\hat{Q}_{11}$	
d=2	n=10	-65	39	-161	121	-120	143	-133	66	-136	195	-126	52
	n=50	-10	5	-25	8	-22	20	-21	6	-37	29	-24	6
	n=100	-7	3	-13	4	-13	10	-8	3	-19	14	-12	3
	n=1000	-1	1	-2	1	-1	1	-1	1	-2	1	-1	1
d=3	n=10	-54	36	-505	560	-123	138	-379	232	-123	138	-379	232
	n=50	-11	5	-140	24	-15	20	-100	16	-15	20	-100	16
	n=100	1	3	-103	11	3	9	-66	7	3	10	-66	7
	n=1000	3	1	-79	1	10	1	-50	1	10	1	-50	1
d=4	n=10	-54	35	-993	1350	-112	133	-697	739	-112	183	620	726
	n=50	-5	5	-238	35	8	16	-170	24	-5	27	-148	21
	n=100	3	3	-188	15	17	8	-122	10	23	12	-104	8
	n=1000	7	1	-149	1	24	1	-91	1	29	1	-74	1
d=5	n=10	-53	46	-1555	4064	-91	138	-1170	3367	-90	157	-1010	1839
	n=50	3	5	-327	66	11	16	-223	43	17	24	-192	40
	n=100	6	2	-255	26	25	8	-163	18	33	11	-149	15
	n=1000	11	2	-201	2	38	1	-127	1	48	1	-103	1

Table 1.1: Bias and variance: figures where multiplied by 1000

Sans surprise, la qualité des résultats s'améliore avec la taille de l'échantillon. Les résultats plus surprenants concernent l'effet de α. Les valeurs plus grandes de α produisent des résultats plus mauvais sur $\boldsymbol{l}$ mais meilleurs sur Q. Pour l'effet de la dimension d, on remarque que la variance de l'estimateur $\hat{\alpha}$ est stable par rapport à d alors que le biais et la variance de $\hat{Q}_{11}$ augmentent avec la dimension. Comme $\hat{Q}_{11}$ est obtenu a partir des $\hat{Q}_{1i}$, on peut justifier l'augmentation du biais et de la variance comme conséquence de l'augmentation du nombre de paramètres. Finalement, on remarque aussi que l'estimateur Q_{11} a un biais négatif et donc par construction, les $\hat{Q}_{1i}$ sont en moyenne positivement biaisés.

Les autres cadres étudiés sont:

- Dans le cas dimension deux, pour un échantillon $\mathrm{HRPar}_{\mathbf{1_d}}(Q, \boldsymbol{l})$ distribué, on considère une structure asymétrique sur $\boldsymbol{l}$, c'est-à-dire $l_1 = -\alpha/2 + \varepsilon$ et $l_2 = -\alpha/2 - \varepsilon$.
- Pour un échantillon dans le domaine d'attraction d'une loi Hüsler-Reiss Pareto, on étudie le biais et la variance de l'estimateur du maximum de vraisemblance.

1.3 Généralités sur le machine learning

Historiquement, la notion de "machine pensante" a été décrite par Turing [175] dans son travail séminal où il présente un célèbre test pour l'intelligence artificielle. Depuis, les avancées dans le domaine de l'intelligence artificielle ont fait d'énormes progrès, dus en particulier aux développements des capacités de calculs. Ainsi, en 1997, Deep Blue a battu Kasparov, alors champion du monde d'échecs. En 2016, Alpha Go, une combinaison entre réseaux de neurones profonds, entraînés par apprentissage supervisé et apprentissage renforcé, et d'arbres de recherche [157] a vaincu Lee Sedol, un des meilleurs joueurs du monde de Go. Jusqu'alors le jeu du Go était considéré comme trop complexe pour que les méthodes brutes surpassent les meilleurs joueurs de Go [102].

D'un autre côté, le monde est plus connecté que jamais depuis l'avènement d'internet (facebook, twitter), et des volumes massifs d'information sont recueillis tout les jours. Devant ce phénomène, de nouveaux problèmes se posent. Des contraintes de temps et de mémoire poussent aux développement de nouvelles méthodes/algorithmes plus rapides pouvant donner en temps limité des solutions partielles. Par exemple, dans le contexte du problème d'estimation de l'inverse parcimonieuse de la matrice de convariance pour des modèles graphiques Gaussien, l'algorithme Graphical-Alternating Minimisation Algorithm (G-AMA), développé par Dalal et Rajaratnam [59] et basé sur l'estimation par maximum de vraisemblance avec une pénalité ℓ^1, propose de maintenir la parcimonie des itérés. Ceci est utile lorsque des contraintes de temps et/ou de dimensions forcent un arrêt prématuré des calculs.

Dans un contexte d'explosion de l'information et des capacités de calcul, l'apprentissage statistique donne une classe d'outils puissants pour le traitement des données massives. Les applications sont nombreuses et diverses et comprennent les systèmes de recommandation [23], la reconnaissance vocale et des formes [83][92], la classification de textes [3], la traduction automatique, etc.

De nombreux problèmes traités peuvent être vus comme des problèmes d'optimisation. Considérons ainsi un simple problème d'apprentissage supervisé. Commençons par quelques définitions qui nous serviront par la suite.
Soit $\mathcal{X}$ l'espace entrée et $\mathcal{Y}$ l'espace réponse.

Definition 1.6. *Un prédicteur f est une fonction mesurable $f : \mathcal{X} \longrightarrow \mathcal{Y}$.*

Definition 1.7. *Une fonction de coût $C : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty)$ est une fonction mesurable telle que $C(y, y) = 0$ et $C(y, y') > 0, y \neq y'$.*

La fonction de coût mesure le coût de prédire y' sachant que la vérité est y. Etant donné un n-échantillon aléatoire $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. sur $\mathcal{X} \times \mathcal{Y}$ de loi jointe $\mathbb{P}_{(X,Y)}$ supposée inconnue et une classe de prédicteur $\mathcal{F}$, pour une fonction de coût C et une réalisation $(x_1, y_1), \ldots, (x_n, y_n)$ du n-échantillon, la plupart des problèmes reviennent à minimiser sur la classe $\mathcal{F}$ le risque empirique défini par

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} C(f(x_i), y_i). \tag{1.32}$$

En d'autres mots, cela revient à trouver le meilleur prédicteur pour la n-réalisation dans la classe $\mathcal{F}$. Des formulations plus théoriques sont toutefois nécessaires pour l'étude des algorithmes. Une formulation plus probabiliste consiste à remplacer le risque empirique $R_n(f)$ par le risque théorique $R(f) = \mathbb{E}[C(f(X_1), Y_1)]$. Bien que minimiser ce risque théorique serait idéal en terme de prédiction, en pratique, le risque théorique est inaccessible car la loi du couple (X_1, Y_1) est inconnue. Si par ailleurs, les prédicteurs $f \in \mathcal{F}$ sont paramétriques de sorte que chaque prédicteur f est uniquement identifié par $\beta \in \mathcal{B}$ ($f = f_\beta$), le problème d'optimisation revient à minimiser $R(f_\beta)$ sur $\mathcal{B}$. Par la suite, on supposera donc que les problèmes sont paramétriques, c'est-à-dire que le prédicteur optimal pour le risque théorique est un prédicteur paramétrique appartenant à un sous-ensemble $\mathcal{F}_\mathcal{B}$ de prédicteurs paramétriques de $\mathcal{F}$. On ne se souciera donc pas du problème de mis-spécification du modèle. Il est aussi fréquent d'introduire une pénalisation comme la pénalité "ridge" ou "lasso" afin de réduire la variance de l'estimateur obtenu.
Par exemple, pour le problème de régression avec $\mathcal{X} = \mathbb{R}^p, \mathcal{Y} = \mathbb{R}, \mathcal{B} = \mathbb{R}^{p+1}$, on définit f_β par

$$f_{\boldsymbol{\beta}}(\mathbf{x}) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j. \tag{1.33}$$

Alors, pour un coût quadratique et un échantillon $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, le problème d'optimisation du risque empirique avec pénalité LASSO s'écrit

$$\min_{\boldsymbol{\beta} \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \|\boldsymbol{\beta}\|_1 \tag{1.34}$$

où λ est un paramètre à régler.
Étant donnée l'importance du LASSO dans le domaine de l'apprentissage, on présentera brièvement quelques propriétés sur l'estimateur LASSO et ses extensions.

L'estimateur lasso

Le lasso est juste une régression linéaire avec une pénalité ℓ^1. Considérons donc le modèle linéaire suivant

$$Y = \mathbf{X}^\top \boldsymbol{\beta} + \varepsilon \tag{1.35}$$

où ε est un terme de bruit. Une telle formulation est pratique car elle nous permet d'éviter le calcul de l'intercept dans le problème d'optimisation et cette formulation peut toujours être obtenue sous réserve de centrer les variables réponses y_i. En statistique, l'usage de la pénalité ℓ^1 a été popularisé par Tibshirani [171] sous le nom de lasso qui a montré sa capacité en tant que sélecteur de variables. Plus précisément, le lasso $\hat{\boldsymbol{\beta}}^{\mathcal{L}}$ est défini par

$$\hat{\boldsymbol{\beta}}^{\mathcal{L}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{1.36}$$

avec $X = (x_{ij})_{i,j}$ et $\mathbf{Y} = (y_i)_i$. On retrouve alors le théorème suivant sur l'erreur quadratique moyenne de l'estimateur lasso:

Theorem 1.13. *Supposons que le bruit ε suit un loi normale de variance σ^2. Supposons que les colonnes de X sont normalisées de telle sorte que $\max_j |\mathbf{X}_j|_2 \leq \sqrt{n}$. Alors l'estimateur du lasso $\hat{\boldsymbol{\beta}}^{\mathcal{L}}$ avec paramètre de régularisation*

$$\lambda = 2\sigma\left(\sqrt{\frac{2\log(2p)}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \tag{1.37}$$

satisfait

$$\mathrm{MSE}(X\hat{\boldsymbol{\beta}}^{\mathcal{L}}) = \frac{1}{n}\|X\hat{\boldsymbol{\beta}}^{\mathcal{L}} - X\boldsymbol{\beta}_*\|_2^2 \leq 4\|\boldsymbol{\beta}_*\|_1\sigma\left(\sqrt{\frac{2\log(2p)}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \tag{1.38}$$

avec probabilité supérieure à $1-\delta$ et $\boldsymbol{\beta}_*$ le vrai paramètre du modèle linéaire.

On peut affaiblir l'hypothèse sur le bruit en supposant que le bruit est sous-gaussien, c'est-à-dire satisfaisant la définition suivante

Definition 1.8. *Une variable aléatoire réelle X est dite sous-gaussienne avec variance proxy σ^2 si $\mathbb{E}[X] = 0$ et sa fonction génératrice des moments vérifie*

$$\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right), \quad s \in \mathbb{R}. \tag{1.39}$$

Un vecteur aléatoire $\mathbf{X} \in \mathbb{R}^p$ est dit sous-gaussien avec variance proxy σ^2 si $\mathbb{E}[\mathbf{X}] = \mathbf{0}$ et $\mathbf{u}^\top\mathbf{X}$ est sous-gaussien avec variance proxy σ^2 pour tout vecteur unitaire $\mathbf{u} \in \mathbb{S}^{p-1}$.

On peut par ailleurs renforcer la condition sur X pour obtenir une meilleure vitesse de convergence. Par exemple, en supposant que X satisfait une propriété d'incohérence, on obtient une vitesse de l'ordre de $\log(2p)/n$. On remarquera par ailleurs que le paramètre de régularisation doit dépendre en théorie du niveau du bruit σ alors qu'en pratique il est inconnu. Une extension du lasso, appelée lasso concomitant introduite par Owen [135] puis étudiée par Sun et Zhang [163].

Definition 1.9. *Soit $\lambda > 0$, l'estimateur du lasso concomitant $\hat{\boldsymbol{\beta}}^{(\boldsymbol{\lambda})}$ est défini par la solution du problème d'optimisation*

$$(\hat{\boldsymbol{\beta}}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \sigma>0}{\operatorname{argmin}} \frac{\|\mathbf{Y} - X\boldsymbol{\beta}\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda\|\boldsymbol{\beta}\|_1. \tag{1.40}$$

Une extension du lasso concomitant pour le cas λ petit a été proposée par Ndiaye, Fercoq, Gramfort et Salmon [128] sous le nom de lasso concomitant lissé avec une étude numérique à l'appui. Une autre approche pour le cas où la variance du bruit est supposée inconnue a été étudiée par Chrétien et Darses [51].

1.3.1 Algorithmes de descente de gradient

Il est commun d'utiliser des algorithmes de type descent de gradient pour optimiser le risque empirique $R_n(f_{\boldsymbol{\beta}})$. Par la suite, pour un risque C et une classe de prédicteur $\{f_{\boldsymbol{\beta}}\}_{\boldsymbol{\beta}\in\mathcal{B}}$, on notera $C(f_{\boldsymbol{\beta}}(X), Y) = G(\boldsymbol{\beta}, X, Y)$. Alors, l'algorithme de descente de gradient (GD) à pas constant consiste à prendre des itérations de la forme

$$\boldsymbol{\beta_{t+1}} = \boldsymbol{\beta_t} - \gamma\frac{1}{n}\sum_{i=1}^{n} \nabla_\beta G(\boldsymbol{\beta_t}, x_i, y_i), \tag{1.41}$$

où $\gamma > 0$ est un paramètre de gain afin de résoudre le problème d'optimisation

$$\min_{\beta\in\mathcal{B}} R_n(f_\beta) := \frac{1}{n}\sum_{i=1}^{n} G(\beta, x_i, y_i). \tag{1.42}$$

Lorsqu'on initialise l'algorithme suffisamment proche de l'optimum et sous certaines hypothèses de régularité sur G (convexité et différentiabilité dans un voisinage de l'optimum), GD converge vers l'optimum à une vitesse linéaire [65]. En remplaçant γ par l'inverse de la hessienne en β_t (ou des approximations), on obtient des algorithmes dits du second ordre. Par exemple, l'algorithme de Newton suivant

$$\boldsymbol{\beta_{t+1}} = \boldsymbol{\beta_t} - \Gamma_t\frac{1}{n}\sum_{i=1}^{n} \nabla G(\boldsymbol{\beta_t}, x_i, y_i), \tag{1.43}$$

avec Γ_t l'inverse de la hessienne en $\boldsymbol{\beta_t}$, est un algorithme du second ordre. Sous certaines hypothèses de régularité de G et sous réserve qu'on initialise l'algorithme suffisamment proche de l'optimum, alors l'algorithme converge vers l'optimum avec une vitesse quadratique. On remarquera néanmoins que chaque itération requiert de calculer n gradient, ce qui est un désavantage en terme de temps de calcul pour n très grand. Par ailleurs, lorsque le nombre de paramètres à estimer est trop grand, inverser la matrice hessienne est aussi coûteux en terme de temps de calcul ou de mémoire.

1.3.2 Généralités sur les algorithmes de gradient stochastique

Par la suite, on va considérer une classe d'algorithme pour résoudre le problème d'optimisation sans pénalité $\min_{\boldsymbol{\beta}\in\mathcal{B}} R(f_{\boldsymbol{\beta}})$.

Les algorithmes de gradient stochastique sous le cadre de Robbins-Monroe

D'un point de vue algorithmique, les algorithmes stochastiques du premier ordre s'avèrent être des méthodes efficaces pour des problèmes en haute dimensionnalité. Si le volume des données n'est pas trop grand, les algorithmes de type "batch" restent des outils utiles pour résoudre le problème d'optimisation. Éventuellement, les algorithmes de type semi-batch où on ne se sert que d'une fraction des données à chaque itération peuvent être utilisés pour accélérer les temps de calcul. Mais pour de grands volumes de données ou des données séquentielles, l'algorithme du gradient stochastique (SGD) introduit par Robbins et Monro [142] est particulièrement populaire. Plus précisément, soit g une fonction C^1 définie sur un ouvert $\mathcal{O}$ de $\mathbb{R}^d$ dans $\mathbb{R}$ dont on cherche les minimums, on cherche les zéros du gradient ∇g de g qui correspondent à des minima. Si $\mathcal{O}$ est convexe et g strictement convexe, le minimiseur est unique et annule le gradient. Alors SGD s'écrit

$$\mathbf{Z_{t+1}} = \mathbf{Z_t} - \eta_t \nabla g(\mathbf{Z_t}) + \eta_t \boldsymbol{\pi_{t+1}} \tag{1.44}$$

avec $Z_0 \in \mathcal{O}$ et $\eta_t \to 0$ une suite de pas décroissante telle que

$$\sum_t \eta_t = \infty, \quad \sum_t \eta_t^2 < \infty \tag{1.45}$$

et $(\boldsymbol{\pi_t})_t$ un terme de perturbation aléatoire indépendant vérifiant $\mathbb{E}(\pi_t) = 0$ et $\mathrm{Var}(\pi_t) < \infty$. Pour faire le lien avec le problème d'optimisation initial, l'idée est que $\nabla g(\mathbf{Z_t}) - \boldsymbol{\pi_{t+1}}$ est une approximation stochastique du gradient de g en $\mathbf{Z_t}$. L'algorithme consiste alors à se déplacer récursivement avec des pas de plus en plus petit vers approximativement le sens opposé de la direction de plus grand accroissement de la fonction g. Finalement, en prenant $g(\boldsymbol{\beta}) = R(f_{\boldsymbol{\beta}}) = \mathbb{E}[G(\boldsymbol{\beta}, X_1, Y_1)]$, on retrouve un algorithme pour résoudre le problème d'optimisation $\min_{\boldsymbol{\beta}} R(f_{\boldsymbol{\beta}})$. Néanmoins, comme la loi jointe est supposée inconnue, la difficulté est le calcul de ∇g. Une façon de contourner le problème est d'intervertir dérivée et intégrale. Dans ce cas, l'approximation stochastique du gradient $\nabla g(\mathbf{Z_t}) - \boldsymbol{\pi_{t+1}}$ est obtenue directement en dérivant G par rapport à $\boldsymbol{\beta}$ au point $\mathbf{Z_t}$. En effet, en posant $\boldsymbol{\pi_{t+1}} = \nabla g(\mathbf{Z_t}) - \nabla G(\mathbf{Z_t}, X_1, Y_1)$ pour tout t, SGD s'écrit

$$\mathbf{Z_{t+1}} = \mathbf{Z_t} - \eta_t \nabla G(\mathbf{Z_t}, X_1, Y_1). \tag{1.46}$$

Etant donné l'échantillon aléatoire $(X_t, Y_t)_t$ avec sa réalisation (x_t, y_t), on retrouve les algorithmes dits online

$$\mathbf{Z_{t+1}} = \mathbf{Z_t} - \eta_t \nabla G(\mathbf{Z_t}, x_t, y_t) \tag{1.47}$$

avec $\mathbf{Z_0} \in \mathcal{O}$. Une telle interversion peut être obtenue en supposant que $G(\cdot, x_1, y_1)$ est convexe et g finie sur un voisinage de $\mathbf{Z_t}$ (cf. Strassen [162]). D'autres méthodes pour pouvoir contourner le calcul de ∇g sont retrouvées dans le livre de Duflo [77]. Comme l'algorithme online ne garde pas en mémoire les observations passées, il est bien adapté pour les cas où les observations sont obtenues progressivement les unes après les autres. Les premiers résultats établissent la convergence en probabilité de Z_t vers l'unique minimiseur de g dans le cas où g est C^2 et fortement convexe [142]. Une amélioration a été donnée par Chung [55].

Theorem 1.14 (Chung)**.** *Suivant les notations précédentes, soit* $\boldsymbol{\beta}_* = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathcal{B}} g(\boldsymbol{\beta})$*, supposons que*

- g *est* C^2 *et fortement convexe,*
- *Pour tout* $\delta > 0$*, il existe une constante positive* $K(\delta)$ *telle que*

$$\inf_{|\mathbf{x}-\boldsymbol{\theta}_*|>\delta} |\nabla g(\mathbf{x})| = K(\delta), \tag{1.48}$$

- *Il existe une constante* K' *tel que pour tout* $\boldsymbol{\beta} \in \mathcal{B}$

$$\mathbb{P}(|G(\boldsymbol{\beta}, X_1, Y_1)| \leq K') = 1. \tag{1.49}$$

Alors pour $a_t = t^{-1-\varepsilon}$*, avec* $(1-C)/2 < \varepsilon < 1/2$*, on a*

$$\mathbb{E}[|\mathbf{Z_t} - \boldsymbol{\beta}_*|^2] \leq \frac{C'}{t^{1-2\varepsilon}} \tag{1.50}$$

où C, C' *sont des constantes positives.*

La normalité asymptotique a été d'abord obtenue, avec des hypothèses supplémentaires, par Chung [55] puis par Sacks [149].

Algorithme SGD pour l'analyse en composante principale (ACP)

Duflo [77] présente un large panorama sur la théorie des algorithmes stochastiques. En particulier, l'exemple du problème de l'analyse en composante principale (ACP) récursive a été étudiée. Comme cet exemple est lié au sujet de la thèse, on présentera brièvement ce problème.

L'analyse en composante principale revient, pour un vecteur aléatoire $\mathbf{Y} \in \mathbb{R}^d$ à variance finie, à faire la décomposition spectrale de la matrice de covariance $\operatorname{var}(\mathbf{Y})$ de $\mathbf{Y}$. Ainsi, en posant

$$\operatorname{var}(\mathbf{Y}) = U\Sigma U^\top \tag{1.51}$$

avec U une matrice orthonormale et Σ une matrice diagonale. Le vecteur $\mathbf{X} = U^\top \mathbf{Y}$ est le vecteur des composantes principales de Y et on retrouve la décomposition

$$\mathbf{Y} = \sum_{i=1}^{d} X_i \mathbf{U_i} \tag{1.52}$$

où les $\mathbf{U_i}$ désignent les colonnes de U et les X_i désignent les composantes de $\mathbf{X}$. L'ACP est souvent vu comme un outil de réduction de la dimension. Pour ce faire, il suffit de tronquer dans la décomposition (1.52) les composantes principales correspondantes aux valeurs propres

les plus petites. En terme de problème d'optimisation, trouver les k composantes principales revient à résoudre le problème d'optimisation suivant

$$\min_{U \in \mathbb{R}^{d \times k}} -\mathrm{tr}(U^\top \mathrm{var}_{\mathbf{Y}}(\mathbf{Y}) U), \quad \text{sous la contrainte } U^\top U = I. \tag{1.53}$$

Néanmoins, comme la loi de $\mathbf{Y}$ est inconnue, la fonction objectif ne peut pas être optimisé. On va par contre supposer qu'on ait accès à une suite de réalisation $(\mathbf{y}_i)_i$ de $\mathbf{Y}$. L'approche classique de l'ACP est de prendre un n-échantillon $\mathbf{y}_1, \ldots, \mathbf{y}_n$ afin d'obtenir le problème empirique

$$\min_{U \in \mathbb{R}^{d \times k}} -\mathrm{tr}(U^\top S_n U), \quad \text{sous la contrainte } U^\top U = I \tag{1.54}$$

avec S_n la matrice de covariance empirique. Il est bien connu que résoudre ce problème revient à faire la décomposition en valeur singulière de la matrice des données centrées. Une approche alternative est de considérer le cas où l'on observe séquentiellement les réalisations de $\mathbf{Y}$. Par exemple, on peut être dans le cas où le n-échantillon est indisponible, que ce soit pour des limitations informatiques (mémoire) ou d'acquisition (données obtenues en continu), ou dans le cas où l'approche empirique n'est pas attractive (en terme de temps de calcul). Étant donné que l'information qui nous intéresse est la matrice de covariance de $\mathbf{Y}$, on supposera par la suite que $\mathbf{Y}$ soit centrée afin d'obtenir des observations non biaisées $\mathbf{y}_i \mathbf{y}_i^\top$ de la covariance. Par ailleurs, sans perte de généralité, on peut supposer que $k = 1$ et on se concentrera donc sur le problème d'optimisation suivant

$$\min_{\mathbf{u} \in \mathbb{R}^d, \|\mathbf{u}\|_2 = 1} -\mathbf{u}^\top \mathrm{var}(\mathbf{Y}) \mathbf{u}. \tag{1.55}$$

En effet, étant donné une solution u_1 au problème (1.55), trouver la seconde composante principale de $\mathbf{Y}$ revient à résoudre

$$\min_{\mathbf{u} \in \mathbb{R}^d, \|\mathbf{u}\|_2 = 1} -\mathbf{u}^\top \left((\mathrm{var}(\mathbf{Y}) - \lambda_1 u_1 u_1^\top \right) \mathbf{u} \tag{1.56}$$

avec λ_1 la valeur propre associée à u_1. Il faut remarquer $\mathrm{var}(\mathbf{Y})$ est symétrique définie positive et la décomposition spectrale de $\mathrm{var}(Y)$ donne

$$\mathrm{var}(Y) - \lambda_1 u_1 u_1^\top = \sum_{\lambda_1 \geq \cdots \geq \lambda_d} \lambda_i u_i u_i^\top - \lambda_1 u_1 u_1^\top. \tag{1.57}$$

Donc la solution de (1.56) est bien la seconde composante principale de $\mathbf{Y}$. Alternativement, il est possible de trouver la seconde composante principale en optimisant une autre variante de (1.55) où on rajoute une contrainte $\langle u_1, u \rangle = 0$. En conclusion, il est possible d'obtenir séquentiellement les composantes principales de $\mathbf{Y}$. Néanmoins, comme le problème (1.55) n'est pas résolvable exactement, calculer successivement les composantes principales induit une erreur à chaque itération.

Pour faire le lien avec l'algorithme SGD, on remarque que la fonction objective est bien convexe mais on a une contrainte qui est en dehors du cadre présenté dans la sous-section

précédente. Néanmoins, SGD (projeté) reste un algorithme viable. Avant de présenter le résultat sur la convergence de SGD projeté, on va étendre le contexte du problème. Soit $A \in \mathbb{R}^{d\times d}$ une matrice semi-définie positive et soit le problème d'optimisation suivant

$$\min_{\mathbf{u}\in\mathbb{R}^d, \|u\|_2=1} -\mathbf{u}^\top A\mathbf{u}. \tag{1.58}$$

Le gradient de la fonction objectif du problème (1.58) au point $\mathbf{u}$ s'écrit $-2A\mathbf{u}$. L'algorithme GD à pas constant pour le problème (1.58) sans contrainte est donné par les itérations

$$\mathbf{Z}_{t+1} = \mathbf{Z}_t + \eta A\mathbf{Z}_t. \tag{1.59}$$

Supposons maintenant que A est inconnu mais qu'on a accès à une suite de matrice aléatoire $(A_t)_t$ i.i.d. semi-définie positive d'espérance A. On a alors accès à un estimateur non biaisé du gradient au point $\mathbf{Z}$ égal à $-2A_t\mathbf{Z}$. L'algorithme SGD à pas décroissant s'écrit alors

$$\mathbf{Z}_{t+1} = \mathbf{Z}_t + \eta_t A_t\mathbf{Z}_t. \tag{1.60}$$

Dans la formulation (1.44), ceci revient à poser $\pi_{t+1} = 2A_t\mathbf{Z}_t - 2A\mathbf{Z}_t$. Finalement, il reste à prendre en compte la contrainte en projetant l'itéré sur l'espace $\{z \in \mathbb{R}^d : \|z\|_2 = 1\}$. Ceci revient à normaliser chaque itéré. On définit ainsi l'algorithme de gradient stochastique projeté à pas décroissant par

$$\mathbf{Z_{t+1}} = \frac{\mathbf{Z_t} + \eta_t A_t\mathbf{Z_t}}{\|\mathbf{Z_t} + \eta_t A_t\mathbf{Z_t}\|_2} \tag{1.61}$$

avec $(\eta_t)_t$ une suite positive telle que $\sum \eta_t = \infty$ et $\sum \eta_t^2 < \infty$. On retrouve alors le théorème [77] suivant sur la convergence de l'algorithme avec un hypothèse de séparation des valeurs propres

Theorem 1.15. *Soit A une matrice réelle $d \times d$ symétrique semi-définie positive ayant pour valeurs propres $\lambda_1 > \lambda_2 \geq \cdots > \lambda_d \geq 0$. Et soit $(A_t)_t$ une suite de matrices aléatoires i.i.d de moyenne A telles que $\|A_t\| \leq K$ (norme opérateur). Soit $\mathbf{Z_0} \in \mathbb{R}^d$ de norme ℓ^2 égal à 1 et $(\eta_t)_t \in \mathbb{R}$ une suite positive telle que $\eta_t < 1/K, t \in \mathbb{N}$ et*

$$\sum_t \eta_t = \infty, \qquad \sum_t \eta_t^2 < \infty. \tag{1.62}$$

Alors $(\mathbf{Z_t})_t$ converge presque sûrement vers le vecteur propre associé à λ_1 (à signe près).

Pour retourner à l'ACP, on remarque que var($\mathbf{Y}$) satisfait les conditions sur la matrice A et que la suite de matrices aléatoires i.i.d. semi-définie positive d'espérance var($\mathbf{Y}$) est donné par $\mathbb{E}(\mathbf{y}_t\mathbf{y}_t^\top$. On peut ainsi construire un algorithme de gradient stochastique projeté pour l'ACP. Plus récemment, Shamir [152][153] a étudié la convergence de SGD pour le problème de l'analyse en composante principale (ACP) sans hypothèse de séparation entre les deux

plus grandes valeurs propres. Ainsi pour l'algorithme du gradient stochastique projeté à pas constant

$$\mathbf{w_{t+1}} = \frac{\mathbf{w_t} + \eta \tilde{A}_t \mathbf{w_t}}{\|\mathbf{w_t} + \eta \tilde{A}_t \mathbf{w_t}\|} \tag{1.63}$$

avec $\tilde{A}_t$ des matrices semi-définies positives i.i.d. vérifiant $\mathbb{E}[\tilde{A}_t] = A$, on a le résultat suivant sur la convergence:

Theorem 1.16. *Sous les hypothèses que*

- *pour un vecteur propre* $\mathbf{v}$ *associé à la plus grande valeur propre de* A, $\frac{1}{\langle \mathbf{v}, \mathbf{w_0} \rangle} \leq p$ *pour un certain* p *positif,*
- *pour* $b \geq 1$, *on a* $\frac{\|\tilde{A}_t\|}{\|A\|} \leq b$ *et* $\frac{\|\tilde{A}_t - A\|}{\|A\|} \leq b$ *(norme opérateur) avec probabilité* 1,

alors après T *itérations avec* $\eta = \frac{1}{b\sqrt{pT}}$ *et avec une probabilité au moins* $\frac{1}{cp}$, $\mathbf{w}_T$ *vérifie*

$$1 - \frac{\mathbf{w_T}^\top A \mathbf{w}}{\|A\|} \leq c' \frac{\log(T) b \sqrt{p}}{\sqrt{T}} \tag{1.64}$$

avec c, c' *constantes positives.*

Extensions

L'inconvénient de SGD est sa vitesse de convergence sous-linéaire (de l'ordre $O(t^{-1})$ pour une fonction objective convexe et différentiable et avec gradient lipschitzien)[142][131]. Ceci est dû au fait qu'on se sert d'une approximation stochastique du gradient à chaque itération et cela induit une variance que l'on doit réduire en prenant un pas décroissant pour obtenir la convergence. De nombreuses améliorations, prenant en compte les spécificités des problèmes rencontrés, ont été développées. On a, par exemple, les algorithmes de gradient stochastique à variance réduite (SVRG) [115] qui consistent à garder en mémoire un estimateur $\overline{\beta}$ obtenu après un certain nombre d'itérations (après chaque m itérations par exemple) puis on se sert de cet estimateur pour corriger l'approximation du gradient. Ce qui, pour un n-échantillon $(x_1, y_1), \ldots, (x_n, y_n)$ s'écrit

$$\boldsymbol{\beta_{t+1}} = \boldsymbol{\beta_t} - \eta(\nabla G(\boldsymbol{\beta_t}, x_t, y_t) - \nabla G(\overline{\boldsymbol{\beta}}, x_t, y_t) + \overline{\boldsymbol{\mu}}) \tag{1.65}$$

où $\overline{\boldsymbol{\mu}} = 1/n \sum_{t=1}^{n} \nabla G(\overline{\beta}, x_t, y_t)$ et $\eta < 1/L$. Il est alors montré, dans le cas où G est fortement convexe, la convergence en espérance à une vitesse géométrique.

Theorem 1.17 (Johnson-Zhang)**.** *Considérons SVRG où à chaque mise-à-jour de* $\overline{\boldsymbol{\beta}}$ *est pris aléatoirement parmi les* $m-1$ *itérés précédant la mise-à-jour. Notons*

$$g_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} G(\cdot, x_i, y_i).$$

Supposons que les fonctions $G(\cdot, x, y)$ sont convexes et $\mathcal{C}^\infty$, et que

$$G(\boldsymbol{\beta}, x_i, y_i) - G(\boldsymbol{\beta'}, x_i, y_i) - \nabla G(\boldsymbol{\beta'}, x_i, y_i)^\top (\boldsymbol{\beta} - \boldsymbol{\beta'}) \leq \frac{L}{2} \|\boldsymbol{\beta} - \boldsymbol{\beta'}\|_2. \tag{1.66}$$

Supposons en plus que la moyenne g_n est fortement convexe, c'est-à-dire

$$g_n(\beta) - g_n(\beta') - \gamma \|\beta - \beta'\|_2^2 \geq \nabla g_n(\beta')^\top (\beta - \beta') \tag{1.67}$$

avec $L \geq \gamma > 0$ pour tout $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Posons

$$\boldsymbol{\beta_*} \in \boldsymbol{\mathcal{B}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} G(\boldsymbol{\beta}, x_i, y_i).$$

Supposons par ailleurs que m est suffisamment grand et que η soit tel que

$$\alpha = \frac{1}{\gamma\eta(1 - 2L\eta)m} + \frac{2L\eta}{1 - 2L\eta} < 1 \tag{1.68}$$

soit vraie. Alors on a la convergence en espérance à vitesse géométrique

$$\mathbb{E}[g_n(\boldsymbol{\beta_t}) - g_n(\boldsymbol{\beta_*})] \leq \alpha^t \; [g_n(\boldsymbol{\beta_0}) - g_n(\boldsymbol{\beta_*})]. \tag{1.69}$$

Par exemple, de Sa, Olukotun et Ré [62] proposent un algorithme de gradient stochastique sur la variété de Stiefel dans lequel la mise à jour se fait le long des géodésiques de la variété.

1.3.3 Généralités sur l'acquisition comprimée

Par leurs travaux précurseurs, Nyquist [134], Shannon [154] et Whittaker [183] ont montré qu'il était possible de reconstruire exactement un signal à temps continu et avec une plage de fréquence finie à partir d'un échantillon (du signal) prélevé à la fréquence de Nyquist, c'est-à-dire au minimum deux fois la plus haute fréquence présente dans le signal. Suivant ces résultats et profitant du développement du numérique et de l'informatique, la quantité de donnée produite a subi une explosion massive. Néanmoins, il arrive souvent que la fréquence de Nyquist soit trop élevée en pratique, et en conséquence on se retrouve avec un échantillon beaucoup trop grand et donc un objet en haute dimension, ou encore, que le coût associé à une telle acquisition soit trop élevé [181].

Modèles parcimonieux

Devant ces difficultés, une idée simple serait de trouver une approximation en plus petite dimension du signal de départ qu'on puisse traiter ou stocker efficacement. Mais avant de continuer, on va définir la notion de parcimonie.

Definition 1.10. *Un vecteur $\mathbf{x} \in \mathbb{R}^d$ est dit k-parcimonieux s'il a au plus k coefficients non nuls, c'est-à-dire $\|\mathbf{x}\|_0 \leq k$ où $\|\mathbf{x}\|_0 = \sum_{i=1}^d 1_{\{x_i \neq 0\}}$.*
On notera Σ_k le sous-ensemble des vecteurs k-parcimonieux de $\mathbb{R}^d$.

Par extension, si $\mathbf{x} \in \mathbb{R}^d$ admet une représentation exacte k-parcimonieuse dans une certaine base alors on va dit que $\mathbf{x}$ est aussi k-parcimonieux.
Grossièrement, si on identifie un signal à temps continu s avec sa représentation digitale $\mathbf{x} \in \mathbb{R}^d$ alors on cherche une approximation k-parcimonieuse $\hat{\mathbf{x}}$ de $\mathbf{x}$ (dans une certaine base). On peut quantifier l'erreur en faisant la meilleure approximation k-parcimonieuse possible par

$$\sigma_k(\mathbf{x}) = \min_{\hat{\mathbf{x}} \in \Sigma_k} \|\mathbf{x} - \hat{\mathbf{x}}\|_1. \tag{1.70}$$

D'une certaine façon, l'idée ressemble beaucoup au cadre de l'ACP, les algorithmes de compression (JPEG, JPG, MPEG, MP3, etc) et de débruitage (décomposition en base d'ondelettes [124]) où on décompose signaux/images dans des bases bien choisies puis on tronque de sorte à obtenir un signal parcimonieux et un dictionnaire (la base) qui permet de reconstruire une approximation de l'image/signal de départ. Néanmoins, ceci n'est pas le but de l'acquisition comprimée. Son but est de transférer la conversion de la représentation du signal en haute dimension vers une représentation parcimonieuse à l'étape même de l'acquisition du signal. Ce qui se résume à passer de

$$\text{signal } s \xrightarrow{\text{acquisition}} \text{représentation digitale } \mathbf{x} \xrightarrow{\text{réd. dim.}} \text{approximation } A\mathbf{x} \tag{1.71}$$

à

$$\text{signal } s \xrightarrow{\text{acquisition}} A\mathbf{x} \tag{1.72}$$

avec $\mathbf{x} \in \mathbb{R}^d, A \in \mathbb{R}^{d' \times d}, d' << d$.
Les travaux fondateurs sont dus à Candès, Romberg et Tao [39][37][38] et Donoho [71], qui ont montré qu'il est possible de reconstruire exactement un signal parcimonieux avec un nombre limité de mesures. Pour formuler le problème de reconstruction, posons

$$\mathbf{y} = A\mathbf{x} \tag{1.73}$$

ou alternativement

$$\mathbf{y} = A\mathbf{x} + \mathbf{e} \tag{1.74}$$

où $\mathbf{x} \in \mathbb{R}^d$ est interprété comme l'échantillon de Nyquist du signal de départ, $A \in \mathbb{R}^{d' \times d}$ est une matrice qui réduit la dimension du signal de sorte que y sont les mesures obtenues par la méthode d'acquisition et $\mathbf{e} \in \mathbb{R}^{d'}$ est un terme de bruit. La question naturelle est alors, quelles sont les conditions nécessaires ou suffisantes sur A de sorte qu'on puisse retrouver ou approximer $\mathbf{x}$ à partir de $\mathbf{y}$. Candès et Tao [39] avaient introduit la notion de propriété d'isométrie restreinte (RIP) pour répondre à cette question.

Definition 1.11 (RIP)**.** *Une matrice $A \in \mathbb{R}^{d' \times d}$ satisfait la propriété d'isométrie restreinte d'ordre k avec constante $\delta_k \in (0, 1)$ si*

$$(1 - \delta_k)\|\mathbf{x}\|_2^2 \leq \|A\mathbf{x}\|_2^2 \leq (1 + \delta_k)\|\mathbf{x}\|_2^2 \tag{1.75}$$

pour tout $\mathbf{x}$ k-parcimonieux.

Mais la condition nécessaire et suffisante pour la reconstruction exacte de tout vecteur k-parcimonieux a été présentée par Cohen, Dahmen et DeVore [71] avec une condition sur le noyau de A.

Definition 1.12 (NSP)**.** *Une matrice $A \in \mathbb{R}^{d \times d'}$ satisfait la propriété NSP d'ordre k avec constante C si*

$$\|\mathbf{h}_T\|_2 \leq C \frac{\|\mathbf{h}_{T^c}\|_1}{\sqrt{k}} \tag{1.76}$$

pour tout $\mathbf{h} \in \mathrm{Ker}(A)$ et tout $T \subset \{0, \ldots, d'\}$ tel que $|T| \leq k$.

La propriété NSP peut être interprétée de plusieurs manières. Une façon d'interpréter NSP est de voir que si $\mathbf{h}$ est k-parcimonieux alors il existe un sous-ensemble T de $\{0, \ldots, d'\}$ tel que $\mathbf{h}_{T^c}$ est le vecteur nul, NSP implique alors que $\mathbf{h}_T$ est aussi le vecteur nul et donc $\mathbf{h}$ est nul. Une matrice ne peut donc pas contenir de vecteur k-parcimonieux dans son noyau. Une matrice qui satisfait la propriété RIP satisfait aussi la propriété NSP [80].

Theorem 1.18. *Soit $A \in \mathbb{R}^{d' \times d}$ une matrice ayant la propriété RIP d'ordre $2k$ avec $\delta_{2k} < 1$. Alors A satisfait la condition NSP d'ordre k avec constante*

$$C = 1 + \frac{1 + \delta_{2k}}{1 - \delta_{2k}}. \tag{1.77}$$

Etant donné un décodeur $\Delta : \mathbb{R}^{d'} \to \mathbb{R}^d$, alors le théorème suivant relie la capacité de reconstruction du signal avec NSP

Theorem 1.19 (Cohen-Dahmen-DeVore)**.** *Soit $A : \mathbb{R}^{d' \times d}$ une matrice d'acquisition et $\Delta : \mathbb{R}^{d'} \to \mathbb{R}^d$ un algorithme de reconstruction tel que*

$$\|\mathbf{x} - \Delta(A\mathbf{x})\| \leq C \frac{\sigma_k(\mathbf{x})}{\sqrt{k}}, \quad \mathbf{x} \in \mathbb{R}^d \tag{1.78}$$

alors A satisfait la propriété NSP d'ordre $2k$.

Ce théorème implique la nécessité de la propriété NSP dans le cadre de la reconstruction exacte d'un signal k-parcimonieux ($\sigma_k(x) = 0$) mais aussi dans le cadre de reconstruction approchée d'un signal quelconque. Néanmoins, pour montrer qu'une matrice $A \in \mathbb{R}^{d' \times d}$ vérifie la propriété NSP ou RIP d'ordre k, il faudrait vérifier les inégalités pour toutes les sous-matrices formées par k colonnes de A. Pour des raisons de temps de calcul, il est parfois préférable de vérifier une propriété plus simple comme la cohérence [73].

Definition 1.13 (Cohérence)**.** *La cohérence $\mu(X)$ d'une matrice $X \in \mathbb{R}^{d' \times d}$ est définie par*

$$\mu(X) = \max_{j \neq j'} \frac{|\langle \mathbf{X_j}, \mathbf{X_{j'}} \rangle|}{\|\mathbf{X_j}\|_2 \|\mathbf{X_{j'}}\|_2} \tag{1.79}$$

Le lien entre RIP et la cohérence d'une matrice A est obtenue par le théorème du cercle de Gershgorin [178].

Lemma 1.1. *Soit $A \in \mathbb{R}^{d' \times d}$ avec colonnes normalisés et cohérence $\mu(A)$. Alors A satisfait la propriété RIP d'ordre k avec $\delta_k = (k-1)\mu$ pour tout $k < 1/\mu$.*

On voit facilement que notre intérêt se porte sur les matrices à cohérence faible.

Reconstruction par le lasso

Après avoir établi les conditions sur la matrice d'acquisition, on peut s'intéresser au problème de reconstruction en lui-même. Une approche naturelle est de considérer le problème d'optimisation

$$\min_{\hat{\mathbf{x}}\in\mathbb{R}^d} \|\hat{\mathbf{x}}\|_0, \qquad \text{tel que } A\hat{\mathbf{x}} = \mathbf{y} \tag{1.80}$$

dans le cas non bruité et

$$\min_{\hat{\mathbf{x}}\in\mathbb{R}^d} \|\hat{\mathbf{x}}\|_0, \qquad \text{tel que } \|A\hat{\mathbf{x}} - \mathbf{y}\|_2 \leq \varepsilon \tag{1.81}$$

pour un certain $\varepsilon > 0$ dans le cas bruité. Néanmoins, de tels problèmes sont connus pour être NP-dur [75][93].
L'approche de Candès, Rombert et Tao est de considérer le problème dit *basic pursuit* [48] pour le problème, c'est-à-dire en remplaçant la "norme" ℓ^0 par la norme ℓ^1, ce qui revient à optimiser

$$\min_{\hat{\mathbf{x}}\in\mathbb{R}^{d'}} \|\hat{\mathbf{x}}\|_1, \qquad \text{tel que } A\hat{\mathbf{x}} = \mathbf{y} \tag{1.82}$$

ou

$$\min_{\hat{\mathbf{x}}\in\mathbb{R}^{d'}} \|\hat{\mathbf{x}}\|_1, \qquad \text{tel que } \|A\hat{\mathbf{x}} - \mathbf{y}\|_2 \leq \varepsilon. \tag{1.83}$$

Candès [35] a montré une borne sur la qualité de l'approximation obtenue par cette méthode

Theorem 1.20 (Candès)**.** *Supposons que A soit RIP d'ordre $2k$ avec $\delta_{2k} = \sqrt{2} - 1$ et $\mathbf{y} = A\mathbf{x} + \mathbf{e}$ avec $\|\mathbf{e}\|_2 \leq \varepsilon$. Alors la solution $\hat{\mathbf{x}}$ du problème* (1.83) *satisfait*

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq C_0 \frac{\sigma_k(\mathbf{x})}{\sqrt{k}} + C_1\varepsilon \tag{1.84}$$

avec

$$C_0 = 2\frac{1-(1-\sqrt{2})\delta_{2k}}{1-(1+\sqrt{2})\delta_{2k}} \qquad \textit{et} \qquad C_1 = 4\frac{\sqrt{1+\delta_{2k}}}{1-(1-\sqrt{2})\delta_{2k}}. \tag{1.85}$$

On peut énoncer un théorème similaire dans le cas non bruité [35].

Theorem 1.21 (Candès)**.** *Supposons que A soit RIP d'ordre $2k$ avec $\delta_{2k} < \sqrt{2} - 1$ et $\mathbf{y} = A\mathbf{x}$. Alors la solution $\hat{\mathbf{x}}$ du problème* (1.82) *satisfait*

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq C_0 \frac{\sigma_k(\mathbf{x})}{\sqrt{k}}. \tag{1.86}$$

Un résultat similaire en prenant NSP au lieu de RIP montre que la propriété NSP est suffisante pour le décodeur *basic pursuit*. Pour compléter le tableau, on présente un théorème pour la propriété de cohérence [74].

Theorem 1.22 (Donoho)**.** *Supposons que A a une cohérence μ et que le signal $\mathbf{x}$ est k-parcimonieux avec $k < (1/\mu + 1)/4$. Par ailleurs supposons que $\mathbf{y} = A\mathbf{x} + \mathbf{e}$. Alors la solution $\hat{\mathbf{x}}$ du problème* (1.83) *satisfait*

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq \frac{\|\mathbf{e}\|_2 + \varepsilon}{\sqrt{1 - \mu(4k-1)}}. \tag{1.87}$$

Bien entendu, les théorèmes présentés ci-dessus ne sont pas exhaustifs, il existe d'autres formulations avec des constantes différentes.

Pour terminer cette introduction, on fait un retour sur le lasso. La théorie de l'analyse convexe [144] nous dit que le problème de programmation linéaire à contrainte conique (1.83) est fortement lié au problème convexe

$$\min_{\hat{\mathbf{x}} \in \mathbb{R}^{d'}} \frac{1}{2}\|A\hat{\mathbf{x}} - \mathbf{y}\|_2^2 + \lambda\|\hat{\mathbf{x}}\|_1 \tag{1.88}$$

En particulier, pour un certain λ, les solutions des deux problèmes coincident. On remarquera alors que la solution de (1.88) correspond à l'estimateur du LASSO [172].

Design de matrices d'acquisition

Une fois des méthodes de reconstruction établies, on peut se poser des questions plus pratiques sur le design de matrices bien conditionnées selon le critère RIP ou de la cohérence. Une approche naturelle est de considérer des matrices déterministes de type Vandermonde par exemple. Mais d'autres approches aléatoires comme les matrices gaussiennes offrent d'autres possibilités. Ainsi, Baraniuk [12] montre des arguments de concentration de la mesure qui impliquent la propriété RIP avec forte probabilité pour des matrices aléatoires. En particulier, les matrices gaussiennes vérifient les conditions nécessaires de concentration.

Theorem 1.23 (Baraniuk)**.** *Soient $(\Omega, \mathcal{A}, \mathcal{P})$ un espace probabilisé et $A \in \mathbb{R}^{d' \times d}$ une matrice aléatoire dont les composantes A_{ij} sont i.i.d. Si, pour tout $\mathbf{x} \in \mathbb{R}^d$, la variable aléatoire $\|A\mathbf{x}\|_2^2$ satisfait, pour tout $\varepsilon \in (0,1)$*

$$\mathbb{P}\left(\frac{|\|A\mathbf{x}\|_2^2 - \mathbb{E}[\|A\mathbf{x}\|_2^2]|}{\mathbb{E}[\|A\mathbf{x}\|_2^2]} \geq \varepsilon\right) \leq 2e^{-d'c(\varepsilon)} \tag{1.89}$$

où $c(\varepsilon)$ est une constante qui ne dépend que de ε alors A a la propriété RIP d'ordre k et $\delta_k \in (0,1)$ avec probabilité supérieure à

$$1 - 2\left(\frac{12}{\delta_k}\right)^k \exp(-c_0(\delta_k/2)n). \tag{1.90}$$

En particulier les matrices gaussiennes avec composantes $\sim \mathcal{N}(0, 1/\sqrt{d'})$ satisfont l'hypothèse avec constante $c(\varepsilon) = \varepsilon^2/4 - \varepsilon^3/6$.

À partir d'une matrice A, on peut aussi considérer des algorithmes d'extraction de sous-matrices bien conditionnées dans le sens cohérence ou RIP. À ce sujet, la littérature se concentre plutôt sur les propriétés d'inversibilité (le problème d'inversibilité restreinte [187]) et de représentation dans le sens où l'image de la sous-matrice approche l'image de A [173].

1.4 Résultats obtenus dans la partie II

La seconde partie regroupe des travaux effectués sous la direction de Stéphane Chrétien.
Notation vectorielle pour la seconde partie: on note $\mathbf{e}$ le vecteur $(1, \dots, 1)$ avec toute les composantes égales à 1. $(e_i)_i$ désigne la base canonique de $\mathbb{R}^d$. Pour une matrice X, $\mathbf{X}_j$ est le vecteur donné par la j-ème colonne de X, $\|X\|$ donne la norme opérateur de X.

1.4.1 Chapitre 5: Feature selection in weakly coherent matrices

Ce chapitre a été tiré d'un manuscript [54] accepté pour publication dans les proceedings au LNCS de la conférence LVA ICA 2018.

Dans ce chapitre, on donne une borne sur la valeur singulière minimale après l'ajout d'une colonne à une matrice potentiellement incohérente. On propose ensuite un algorithme de sélection de colonne à partir de cette borne.

Dans le contexte du problème de sélection de colonne, pour une matrice $X \in \mathbb{R}^{n\times p}$, on cherche à trouver une sous-matrice $X_T, |T| = t$ de X ayant de bonnes propriétés spectrales [173]. Dans notre cas on cherche à ce que les valeurs singulières $\lambda_i, 1 \leq i \leq t$ de X_T vérifient

$$|\lambda_i| \geq C, \quad C > 0, 1 \leq i \leq t. \tag{1.91}$$

Ceci revient à trouver X_T tel que les valeurs propres non nulles de $X_T X_T^\top$ soient strictement plus grandes qu'un certain seuil strictement positif K.

Par la suite, on considère des matrices dont les colonnes sont normalisées. Notre résultat principal est le suivant:

Theorem 1.24. *Soit $T_0 \subset \{1, \dots, p\}$ avec X_{T_0} une sous-matrice de X. Soit $\lambda_1(X_{T_0}X_{T_0}^\top) \geq \dots \geq \lambda_{s_0}(X_{T_0}X_{T_0}^\top)$ les valeurs propres de $X_{T_0}X_{T_0}^\top$. Alors, pour $X_j \in \mathbb{R}^n$, on a*

$$\lambda_{s_0+1}(X_{T_0}X_{T_0}^\top + \mathbf{X_j}\mathbf{X_j}^\top) \geq \lambda_{s_0}(X_{T_0}X_{T_0}^\top) - \min\left(\|X_{T_0}^\top \mathbf{X_j}\|_2, \frac{\|X_{T_0}^\top \mathbf{X_j}\|_2^2}{1 - \lambda_{s_0}(X_{T_0}X_{T_0}^\top)}\right). \tag{1.92}$$

La preuve se base sur le lemme d'entrelacement de Cauchy. Ainsi les valeurs propres de $X_{T_0}X_{T_0}^\top + \mathbf{X_j}\mathbf{X_j}^\top$ sont entrelacées avec celles de $X_{T_0}X_{T_0}^\top$. La plus petite valeur propre non nulle de $X_{T_0}X_{T_0}^\top + \mathbf{X_j}\mathbf{X_j}^\top$ est alors la plus petite racine de

$$f(x) = 1 - \sum_{i=1}^{n} \frac{\langle \mathbf{X_j}, \mathbf{u_i}\rangle^2}{x - \lambda_i(X_{T_0}X_{T_0}^\top)}. \tag{1.93}$$

Comme cette fonction est croissante sur $(0, \lambda_{s_0}(X_{T_0}X_{T_0}^\top))$, on trouve alors une fonction majorante à f dont on connaît la racine sur l'intervalle qui permet de conclure.

En pratique, les sous-matrices de X_T ont une meilleure cohérence que X qu'on peut quantifier par un facteur $\alpha \in (0, 1]$. On peut alors réécrire le théorème précédent en terme du paramètre α.

Corollary 1.1. *Soit $T_0 \subset \{1, \ldots, p\}$ avec X_{T_0} une sous-matrice de X. Et supposons que*

$$\|X_{T_0}^\top \mathbf{X_j}\|_2^2 \leq \alpha s_0 \mu(X)^2.$$

Alors

$$\lambda_{s_0+1}\left(X_{T_0}X_{T_0}^\top + \mathbf{X_j}\mathbf{X_j}^\top\right) \geq \lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right) - \min\left(\sqrt{\alpha s_0 \mu^2}, \frac{\alpha s_0 \mu^2}{1 - \lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right)}\right). \tag{1.94}$$

Bien entendu, on peut répéter l'opération pour trouver une borne inf pour la valeur singulière minimale après l'ajout de n colonnes.

Ce résultat suggère un algorithme "greedy" pour la sélection de colonne. L'idée est simple: pour obtenir une sous-matrice X_T de X, on sélectionne les colonnes une à une. La colonne choisie est celle minimisant la norme du produit scalaire avec les colonnes déjà sélectionnées. La condition d'arrêt est alors donnée par la borne inférieure obtenue par le théorème qui sert de garantie sur l'inversibilité de la sous-matrice obtenue. D'où l'algorithme suivant :

Input: a matrix $X \in \mathbb{R}^{n\times p}$, $\varepsilon > 0$
Output: a submatrix X_T
Set $s = 1$ and choose a random singleton $T = \{j^{(1)}\} \subset \{1, \ldots, p\}$.
Set $\eta^{(1)} = 1$.
while $\eta^{(s)} \geq 1 - \varepsilon$ **do**
  Set

$$j^{(s)} \in \operatorname{argmin}_{j\in\{1,\ldots,p\}\setminus T} \quad \|X_T^\top \mathbf{X_j}\|_2.$$

  Set

$$\alpha^{(s)} = \|X_T^\top \mathbf{X_{j^{(s)}}}\|_2^2/(s\mu(X)^2).$$

  Set $T = T \cup \{j^{(s)}\}$.
  Set

$$\eta^{(s+1)} = \eta^{(s)} - \min\left(\sqrt{\alpha^{(s)}}\, s\mu, \frac{\alpha^{(s)}\mu(X)^2 s}{1 - \lambda_s(X_T^\top X_T)}\right).$$

  Set $s \leftarrow s + 1$.
end
return X_T.

Algorithm 1: Greedy column selection

Remarquons que l'algorithme requiert le calcul de la valeur propre minimale à chaque itération ainsi que μ. Un autre algorithme peut être obtenu en remplaçant la valeur propre minimale $\lambda_s(X_T^\top X_T)$ par sa borne inf théorique $\eta^{(s)}$ et $\alpha^{(s)}\mu(X)^2 s$ par $\|X_T^\top \mathbf{X}_{j^{(s)}}\|_2^2$.

Pour illuster cet algorithme sur des données réelles, nous considérons le problème d'extraction de séries temporelles "représentatives" d'une grande base de données. Ainsi, on considère un ensemble de 1479 séries temporelles de taille 39 obtenues par transformation non-linéaire de données satellite InSAR pour des soucis d'identifiabilité. À partir

de ces données, nous avons extrait 150 séries temporelles sequentiellement en minimisant $\|X_T^\top \mathbf{X_j}\|_2, j \notin T$ à chaque itération.

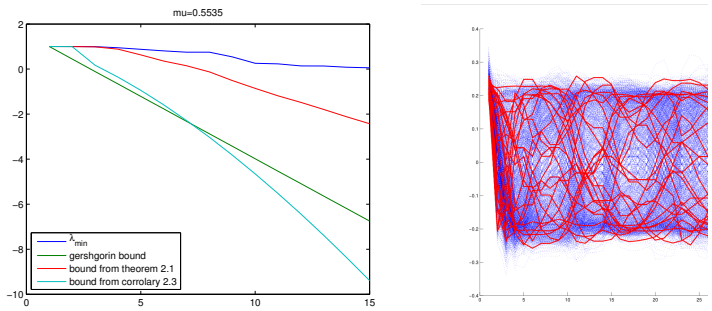


Figure 1.1: Gauchet: Evolution de la plus petite valeur singulière de la sous-matrice obtenue par l'algorithme 1. Droite: Séries temporelles extraites.

La figure 1.1 montre le comportement de l'algorithme en fonction des itérations. On remarquera que dans la cas où la cohérence est forte, les bornes obtenues par le théorème sont meilleures que celle obtenues par le théorème de Gershgorin.

Puis, pour comparer avec les méthodes de sélection de colonne "classique" comme l' algorithme CUR, on a généré 100 matrices avec 100 lignes et 10000 colonnes et on a extrait 10 colonnes avec la méthode CUR et l'algorithme greedy puis on a comparé les valeurs singulières minimales des matrices extraites.

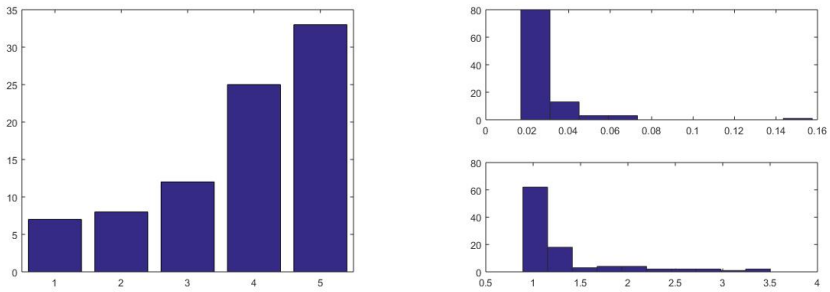


Figure 1.2: Gauche: nombre de valeurs singulières des sous-matrices extraites par l'algorithme 1 plus grande que celle obtenue par la méthode CUR parmis les 5 plus petites valeurs singulières sur 100 expérience Monte-Carlo. Haut-Droit: histograme des temps de calcul pour l'algorithme 1. Bas-Droit: histogramme des temps de calcul pour la méthode CUR [28].

Notre algorithme performe empiriquement mieux que l'algorithme CUR, aussi bien en temps de calcul que sur les valeurs singulières extraites, en effet l'algorithme prend en général moins de $0,5$ seconde alors que CUR prend plus d'une seconde.

1.4.2 Chapitre 6: Small coherence implies the weak Null Space Property

Dans ce chapitre, on montre qu'une faible cohérence implique une version faible de la propriété NSP.

Definition 1.14. *On dit que $X \in \mathbb{R}^{n\times p}$ vérifie weak-NSP(s, C, π) si pour au moins une proportion π des sous-ensembles d'indices $T_0 \subset \{1, \dots, p\}$ tels que $|T_0| = s$ et pour tout $\mathbf{h} \in \mathrm{Ker}(X)$, on a*

$$\|\mathbf{h_{T_0}}\|_2 \leq C\|\mathbf{h_{T_0^c}}\|_1/\sqrt{s}. \tag{1.95}$$

Autrement

$$\frac{\#\left\{T_0 \subset \{1, \dots, p\} : |T_0| = s \ \&\ \forall h \in \mathrm{Ker}(X), \mathbf{h_{T_0^c}}\|_1/\sqrt{s}\right\}}{\#\left\{T_0 \subset \{1, \dots, p\} : |T_0| = s\right\}} \geq \pi. \tag{1.96}$$

Notre résultat principal est le suivant

Theorem 1.25. *Soient $X \in \mathbb{R}^{n\times p}, s_0 \leq n$ et $\alpha > 0$. Soit μ la cohérence de X. Sous les hypothèses que s_0 et μ vérifient*

$$s_0 \leq \frac{1}{16(1+\alpha)e^2}\frac{p}{\|X\|^2 \log p} \tag{1.97}$$

$$\mu \leq \min\left\{(288 s_0^{5/2}(2s_0^{3/2}+1))^{-1/2}, (1.5s_0^4 + 6s_0^{5/2} + 2s_0)^{-1/2}, (4(1+\alpha)\log p)^{-1}\right\} \tag{1.98}$$

Alors la matrice X vérifie la propriété weak-NSP(s_0, C, π) avec $\pi = 1 - 1944/p^\alpha$ et

$$C = \frac{\lambda_1 - \lambda_{s_0} + 3s_0(\varepsilon_{\max} + \varepsilon_{\min})}{\lambda_1 - 3s_0\varepsilon_{\min}} \tag{1.99}$$

avec

$$\varepsilon_{min} = \frac{\frac{1}{4}\ s_0^3\ \mu^2 + s_0^{3/2}\ \mu}{(3 - 4s_0\mu^2)}, \quad \varepsilon_{max} = 144 s_0^3\ \mu^2 + 72 s_0^{3/2}\mu.$$

La preuve repose sur des résultats de perturbations de valeurs propres. Plus concrètement, on montre que si on a $T_0 \subset \{1, \dots, p\}, T_1 \subset \{1, \dots, p\}$ avec $|T_0| = s_0, |T_1| = 3s_0, T_1 \cap T_0$ et $\lambda_1 \geq \dots \geq \lambda_{s_0}$ les valeurs propres non nulles de $X_{T_0}X_{T_0}^\top$, alors on peut encadrer les valeurs propres non nulles de $X_{T_0\cup T_1}X_{T_0\cup T_1}^\top$ par une majoration/minoration de λ_1, λ_{s_0} plus un terme de perturbation. Sachant ce résultat, pour l'appliquer, on a besoin d'un sous-ensemble T_0 de $\{1, \dots, p\}$ de cardinal s_0 dont on connaît un encadrement des valeurs propres. Pour cela, on applique un théorème de Chrétien-Darses [52] qui nous assure l'existence avec forte probabilité (π) d'un tel sous-ensemble. Si on trouve un tel sous-ensemble, alors en divisant l'ensemble

$\{1, \ldots, p\} \setminus T_0$ en des sous-ensembles $T_1, \ldots, T_{n_{s_0}}$ de taille s_0 et en prenant $T = T_0 \cup T_1$, on montre que

$$(\lambda_{s_0} - 3\ s_0\ \varepsilon_{min})\ \|\mathbf{h_T}\|_2^2 \leq \|X_T \mathbf{h_T}\|_2^2 \cdot, \tag{1.100}$$

et aussi

$$\|X_T \mathbf{h_T}\|_2^2 \leq (\lambda_1 - \lambda_{s_0} + 3\ s_0\ (\varepsilon_{max} + \varepsilon_{min}))\ \|\mathbf{h_T}\|_2\ \frac{\|\mathbf{h_{T_0^c}}\|_1}{\sqrt{s_0}}. \tag{1.101}$$

Ce qui nous permet de conclure.

1.4.3 Chapitre 7: Incoherent submatrix selection via approximate independence sets in scalar product graphs

Dans ce chapitre, on considère le problème de sélection de sous-matrice incohérente comme un problème de sélection d'un sous-ensemble indépendent maximal d'un graphe. On propose alors un estimateur de type spectral pour ce problème d'ensemble indépendent.

Etant donné une matrice $X \in \mathbb{R}^{n \times p}$ et un seuil η, on cherche la plus grande sous-matrice composée des colonnes de X ayant une cohérence inférieure à η. On associe alors à (X, η) le graphe $\mathcal{G} = (V, E)$ avec

- $V = \{1, \ldots, p\}$,
- $(j, j') \in E$ si et seulement si $|\langle \mathbf{X_j}, \mathbf{X_{j'}} \rangle| > \eta$.

Clairement, on veut sélectionner les noeuds isolés dans le graphe $\mathcal{G}$. Ce qui correspond à trouver les ensembles indépendents du graphe. Un sous-ensemble indépendent du graphe $\mathcal{G} = (V, E)$ est un sous-ensemble de V dont les sommets sont deux à deux non-adjacent. Comme on cherche la plus grande sous-matrice, cela revient alors à chercher un sous-ensemble indépendent de cardinal maximal. En effet, chaque sous-ensemble indépendent maximal correspond à une sous-matrice de X ayant une cohérence inférieure à η de taille maximale. Le problème d'optimisation associé est donné par

$$\max_{\boldsymbol{\rho} \in \{0,1\}^p} \mathbf{e}^\top \boldsymbol{\rho}, \quad \text{tel que } \boldsymbol{\rho}^\top M \boldsymbol{\rho} = 0 \tag{1.102}$$

où M est la matrice d'adjacence (ou de pseudo-adjacence) du graphe $\mathcal{G}$ et $\mathbf{e}$ le vecteur 1_p. En effet, pour tout $\rho = (\rho_1, \ldots, \rho_p) \in \{0,1\}^p$ tel que $\boldsymbol{\rho}^\top M \boldsymbol{\rho} = 0$, l'ensemble $\{j \in \{1, \cdots, p\} : \rho_j = 1\}$ forme une sous-ensemble indépendent de $\mathcal{G}$. On peut donc voir ρ comme un indicateur de sélection de colonne de la matrice X et maximiser $\mathbf{e}^\top \rho$ (qui est aussi égal $\|\rho\|_0$) revient à prendre un maximum de colonnes. Ce problème appartient à la classe des problèmes NP-dur. On propose alors le problème relaxé

$$\max_{\boldsymbol{\rho} \in \{0,1\}^p} \mathbf{e}^\top \boldsymbol{\rho}, \quad \text{tel que } \boldsymbol{\rho}^\top M \boldsymbol{\rho} \leq r. \tag{1.103}$$

Un changement de variable nous permet d'obtenir le problème suivant

$$\max_{\boldsymbol{\rho}\in\{-1,1\}^p} \mathbf{e}^\top \frac{\mathbf{z}+\mathbf{e}}{2}, \quad \text{tel que } \frac{1}{4}(\mathbf{z}+\mathbf{e})^\top M(\mathbf{z}+\mathbf{e}) \leq r \tag{1.104}$$

Finalement, en relaxant la contrainte $\rho \in \{-1, 1\}$ avec ρ sur la sphère, on obtient le problème pénalisé

$$\min_{\|\mathbf{z}\|_2^2=p} -\mathbf{e}^\top \frac{\mathbf{z}+\mathbf{e}}{2} + \lambda\left(\frac{1}{4}(\mathbf{z}+\mathbf{e})^\top M(\mathbf{z}+\mathbf{e}) - r\right) \tag{1.105}$$

avec $\lambda > 0$ un multiplicateur de Lagrange. Notre résultat principal est alors le suivant

Theorem 1.26. *Notons x_2^* la solution du problème d'optimisation*

$$\min_{\|\mathbf{z}\|_2^2=p} -\mathbf{e}^\top \frac{\mathbf{z}+\mathbf{e}}{2} + \lambda\left(\frac{1}{4}(\mathbf{z}+\mathbf{e})^\top M(\mathbf{z}+\mathbf{e}) - r\right). \tag{1.106}$$

et $\boldsymbol{\rho}^*$ une solution du problème

$$\max_{\boldsymbol{\rho}\in\{0,1\}^p} \mathbf{e}^\top \boldsymbol{\rho}, \quad \textit{tel que } \boldsymbol{\rho}^\top M \boldsymbol{\rho} = 0. \tag{1.107}$$

Soit $\delta > 0$ de sorte que $M_\delta = M + \delta I$ soit définie positive. Posons alors λ_1 la plus petite valeur propre de M_δ, $\phi_1, \ldots, \phi_p$ les vecteurs orthonormaux de M_δ et q_1, q_2 définis par

$$\mathbf{q_1} = \frac{1}{\sqrt{p}} M_\delta \mathbf{e}, \quad q_2 = -\frac{1}{\sqrt{p}}\Big(\frac{(1+\delta)}{\lambda}\ \mathbf{e} - M_\delta \mathbf{e}\Big) \tag{1.108}$$

Posons finalement

$$\gamma_{k,i} = \boldsymbol{\phi}_i^\top \mathbf{q}_k \tag{1.109}$$

pour $k = 1, 2$ et $i = 1, \ldots, p$. Alors on a

$$\|\mathbf{x_2}^* - \boldsymbol{\rho}^*\|_\infty \leq \sqrt{p}\left(\frac{(1+\delta)}{\lambda(\lambda_1-\mu_2)} + \frac{\|\gamma_1\|_2\ r^*}{(\lambda_1-\mu_1)(\lambda_1-\mu_2)}\right)^2,$$

avec r^* donné par

$$r^* = (\lambda_p - \mu_1)\phi\left(p\ \frac{\gamma_{1,\max}^2}{\gamma_{1,\min}^2}\frac{\frac{(1+\delta)^2}{\lambda^2} + \frac{2}{\sqrt{p}}\frac{1+\delta}{\lambda}\mathbf{e}^\top M_\delta \mathbf{e}}{2\ \|\gamma_2\|_2^2}\right)$$

et ϕ la fonction inverse $x \mapsto x/(1+x)^3$.

L'idée de la preuve est de montrer que la problème (1.105) peut être réécrit comme un problème de minimisation (P) de la distance à l'oracle $\boldsymbol{\rho}^*$ plus un terme de pénalité linéaire en $1/\lambda$, c'est-à-dire que $\boldsymbol{\rho}^*$ est solution de (P) lorsque $1/\lambda = 0$. On note (P_0) le problème (P) dans le cas $1/\lambda = 0$. On utilise ensuite un résultat de perturbation qui estime la distance entre les solutions de (P) et (P_0) en fonction de $\gamma_1 = (\gamma_{1,i})_i$ et $\gamma_2 = (\gamma_{2,i})_i$. Pour cela, on construit une fonction dont $\mathbf{x_2^*}$ est solution (grossièrement car il s'agit en fait d'un paramètre qui identifie $\mathbf{x_2^*}$) qu'on étudie dans un voisinage de $\boldsymbol{\rho}^*$ (de la même façon, il s'agit en fait d'un paramètre qui identifie $\boldsymbol{\rho}^*$). Le théorème de Neuberger [132] nous assure alors l'existence d'un zéro de notre fonction dans une boule dont on connaît le rayon.

1.4.4 Chapitre 8: Average performance analysis of the projected gradient method for online PCA

Ce chapitre est tiré d'un manuscrit [53] accepté pour publication dans les proceedings au LNCS de la conférence LOD 2018.

Dans ce chapitre, on étudie la version online de l'algorithme du gradient stochastique pour l'analyse en composante principale (ACP). On obtient un contrôle sur la performance de l'algorithme sans hypothèse de séparation entre les deux plus grandes valeurs propres puis on donne une méthode pratique pour gérer le pas de l'algorithme grâce à une version récente [121] de l'algorithme "Hedge" [88]

Le problème qu'on étudie est une version en ligne de l'algorithme de gradient stochastique pour ACP. Dans notre contexte, on observe les entrées de la matrice A une à une. Une façon de voir le problème est de considérer une matrice $A \in \mathbb{R}^{d\times d}$ et $A_1, A_2, \dots$ matrices aléatoires telles que

$$A_t = d^2 A_{I_t,J_t} \mathbf{e_{I_t}} \mathbf{e_{J_t}}^\top \tag{1.110}$$

où (I_t, J_t) est uniformément distribuée sur $\{1, \dots, n\}^2$. L'algorithme de gradient stochastique projeté à pas constant est alors donné par

$$\mathbf{w_{t+1}} = (I + \eta A_t)\mathbf{w_t} / \|(I + \eta A_t)\mathbf{w_t}\|_2 \tag{1.111}$$

avec $\mathbf{w_0} \in \mathbb{R}^d$. Sans perte de généralité, on supposera par la suite que $\|A\| = 1$.

Notre résultat est le suivant

Theorem 1.27. *Soit $\varepsilon > 0$ et supposons que $1/p < \langle \mathbf{w_0}, \mathbf{v} \rangle$ pour un vecteur propre v associé à la plus grande valeur propre de A. Soit V_T défini par*

$$V_T = \mathbf{w_0}^\top \prod_{t=T}^{1} (I + \eta A_t)^* ((1-\varepsilon)I - A) \prod_{t=1}^{\top} (I + \eta A_t)\mathbf{w_0}. \tag{1.112}$$

Alors, pour T satisfaisant

$$T > \max\left(\frac{4p^2 d^2}{\varepsilon}, \frac{\log 4p\varepsilon^{-1}}{\log(1 + \frac{\varepsilon}{pd^2})} \right) \tag{1.113}$$

et $\eta = \frac{\varepsilon}{4pd^2}$, on a

$$\mathbb{E}[V_T] \leq -\frac{\varepsilon}{4p}(1 + 2\eta)^T. \tag{1.114}$$

Une conséquence du théorème est qu'en espérance, après T itérations de l'algorithme, on a $\mathbb{E}[\mathbf{w_T} A \mathbf{w_T}] \geq 1 - \varepsilon$. La preuve repose sur la majoration récursive de l'espérance de l'erreur commise à chaque itération qu'on obtient grâce à la connaissance de la distribution des A_t.

Une question naturelle se pose sur le choix du pas η de l'algorithme qui a un fort impact sur la vitesse. On propose l'algorithme suivant :

Input: The tolerance $\epsilon \in (0,1)$, and the algorithm's parameters $R,\ K \in \mathbb{N}_*$, $\rho \in (0,1)$ and $\beta > 0$.

Output: convergence criterion L

Burn-in period: **while** $\max_{k=1,\dots,K}\ L_t^{(k)} < 1 - 10\ \epsilon$ **do**

For $\eta \in \{\rho^k\}_{k=1:K}$, run R gradient iterations in parallel whose iterates are denoted by $\mathbf{w_t^{(k,r)}}$, $t = 1,\dots,B$. Define $\pi_0^{(k)} = 1/K$, $k = 1,\dots,K$. For $t = 1,\dots,B$, let

$$L_t^{(k)} = \frac{2}{R(R-2)} \sum_{r<r'=2,\dots,R} \langle \mathbf{w_t^{(k,r)}}, \mathbf{w_t^{(k,r')}} \rangle, \tag{1.115}$$

and for $k = 1,\dots,K$, define

$$\pi_{t+1}^{(k)} = \pi_t^{(k)}\ \exp\left(\beta\ L_t^{(k)}\right). \tag{1.116}$$

end

After burn-in:

Reset R to 1 and K to 1.

Normalise π.

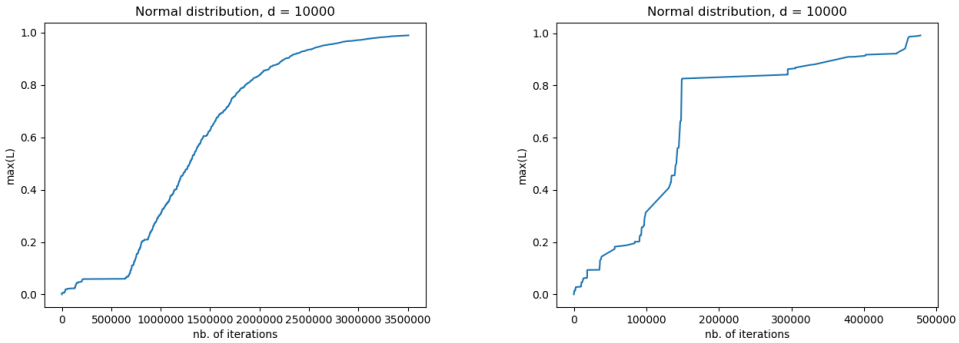
At each step $t = B+1,\dots$, choose the stepsize with probability π_B.

Stop when $\geq 1 - \epsilon$.

return $L_t^{(1)}$.

Cet algorithme consiste à prendre aléatoirement la vitesse de convergence avec des probabilités proportionelles à la "vitesse de convergence moyenne" calculée lors du "burn-in".

Pour illuster la convergence moyenne de notre algorithme ainsi que l'efficacité de notre méthode de sélection de pas de gradient, on propose la simple expérience suivante: on simule une matrice aléatoire gaussienne de taille 1000×10000 que l'on renormalise. On affiche la convergence de l'algorithme de gradient stochastique avec une vitesse arbitraire et on fait de même avec le pas obtenu par la méthode proposée.



(a) $K = 1$ (b) $K = 20$

Figure 1.3: Convergence de $(L_t^{(1)})_{t\in\mathbb{N}}$ en fonction du nombre d'itérations: (a) est pour le cas d'un choix arbitraire de pas égal à 2^{-4} et (b) montre le comportement de la méthode en utilisant la procédure de la Section 8.4.1 pour des valeurs potentielles égales à 2^{-3}, 2^{-2}, 2^{-1}, 1, 2, ..., 2^{17}.

On montre ainsi par cette expérience que l'algorithme de gradient stochastique converge bien en pratique et la sélection du pas par notre algorithme accélère la vitesse de convergence.

Part I

Partie 1

Chapter 2

Simple models for multivariate regular variations

2.1 Introduction

Regular variation is a fundamental notion in extreme value theory that was widely popularised by Resnick [140]. As a simple illustration of the importance a regular variation in univariate extreme value theory, consider an independent and identically distributed (i.i.d.) sequence $X, X_1, X_2, \ldots$ of positive random variables with cumulative distribution F. For $n \geq 1$, let $a_n = F^{\leftarrow}(1 - 1/n)$ be the quantile of order $1 - 1/n$ of F. Then the following statements are equivalent:

i) the tail function $1 - F$ is regularly varying at infinity with index $-\alpha < 0$, i.e.,

$$\lim_{u \to \infty} \frac{1 - F(ux)}{1 - F(u)} = x^{-\alpha}, \quad x > 0;$$

ii) the rescaled maximum $a_n^{-1} \max(X_1, \ldots, X_n)$ converges in distribution as $n \to \infty$ to a standard α-Fréchet distribution, i.e.,

$$\lim_{n \to \infty} \mathbb{P}\left[a_n^{-1} \max(X_1, \ldots, X_n) \leq x\right] = \exp(-x^{-\alpha}), \quad x > 0;$$

iii) the rescaled exceedance $u^{-1}X$ of X given $X > u$ converges in distribution as $u \to \infty$ to a standard α-Pareto distribution; i.e.,

$$\lim_{u \to \infty} \mathbb{P}\left[u^{-1}X > x \mid X > u\right] = x^{-\alpha}, \quad x > 1.$$

iv) the sample point process $\{X_i/a_n,\ 1 \leq i \leq n\}$ converges to a Poisson point process on $(0, \infty)$ with intensity $\alpha x^{-\alpha-1} \mathrm{d}x$.

The equivalence i)-ii) dates back to Gnedenko [91], the equivalence ii)-iii) is due to Balkema and de Haan [7] and the equivalence i)-iv) can be found in Resnick [140]. As will be reviewed in Section 2.2.1, a similar result holds in the multivariate setting and multivariate regular variations is crucial in multivariate extreme value theory.

Historically, multivariate extreme value theory has been developed by considerations on the asymptotic behaviour of i.i.d. random vectors. Key early contributions are the papers by Tiago de Oliveira [168], Sibuya [156], de Haan and Resnick [98], Deheuvels [63]. The general structure of multivariate extreme value distribution has been characterised by de Haan and Resnick [98] in terms of the so-called spectral representation. Domain of attractions have been characterised by Deheuvels [63] that pointed out the convergence of the dependence structure to an extreme value copula. Since then a rich literature has emerged on modelling or statistical aspects of the theory, of which a nice recent review from the copula viewpoint is provided by Gudendorf and Segers [94].

More recent developments focus on exceedances over high threshold in a multivariate setting and the so called multivariate generalised Pareto distributions. Seminal papers in that direction are Coles and Tawn [58] and Rootzen and Tajvidi [148]. Further recent development on modelling and statistical aspects include Rootzen et al. [147] and Kiriliouk et al. [116].

In this framework, the motivations of the present chapter are twofold. In a first part corresponding to Section 2, we revisit multivariate extreme value theory models and put the emphasis on regular variations and the limiting homogeneous measure. More precisely, a multivariate extension of the celebrated Breiman Lemma due to Davis and Mikosch [61] allows us to construct a regularly varying random vectors as a product of a heavy tailed random variable (thought of as a radial component) and a sufficiently integrable random vector (thought as a spectral component). The limiting homogeneous measure is easily characterised and, for specific choice of the spectral component, we recover standard parametric models from multivariate extreme value theory such as the Hüsler-Reiss [110], extremal-t [133], logistic, negative logistic or Dirichlet models [58]. We believe putting the emphasis on the exponent measure is important since it is the fundamental notion that unifies maxima, exceedances or point processes approaches in extreme value theory. On the other hand, from the copula point of view, the multivariate Breiman Lemma provides a general framework for deriving extreme value copula models closely related to the results by Nikoloulopoulos et al. [133] or Belzile and Nešlehová [20].

2.2 A simple model for multivariate regular variations

2.2.1 Preliminaries on multivariate regular variations

Following Hult and Lindskog [107], we define multivariate regular variation in terms of M_0-convergence in $\mathbb{R}^d$ rather than vague convergence in $[-\infty, \infty]^d \setminus \{\mathbf{0}\}$. This is completely equivalent in the multivariate setting but M_0-convergence can be more easily generalised to a metric space.

Consider the space $M_0(\mathbb{R}^d)$ of Borel measures μ on $\mathbb{R}^d \setminus \{\mathbf{0}\}$ that assigns finite mass on sets bounded away from 0, that is $\mu(\mathbb{R}^d \setminus O)$ is finite for all O open neighborhood of $\mathbf{0}$. A sequence $\mu_n \in M_0(\mathbb{R}^d)$ is said to converge to $\mu \in M_0(\mathbb{R}^d)$, noted $\mu_n \overset{M_0}{\to} \mu$, if $\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu$ for all bounded continuous function f that vanishes on a neighbourhood of 0.

A random vector $\mathbf{X}$ on $\mathbb{R}^d$ is called regularly varying with sequence $a_n \to +\infty$ if

$$n\mathbb{P}(\mathbf{X}/a_n \in \cdot) \overset{M_0}{\to} \Lambda \quad \text{as } n \to \infty$$

with a non-zero limit measure $\Lambda \in M_0(\mathbb{R}^d)$. Necessarily, there exists $\alpha > 0$, called the tail index of $\mathbf{X}$, such that the limit measure is homogeneous of order α, i.e.,

$$\Lambda(uA) = u^{-\alpha}\Lambda(A) \quad u > 0, A \subset \mathbb{R}^d \setminus \{\mathbf{0}\} \text{ Borel.}$$

Furthermore, the sequence (a_n) is regularly varying at infinity with index $1/\alpha$ and a possible choice for the normalising sequence a_n is

$$a_n = \inf\{x > 0;\ \mathbb{P}(\|\mathbf{X}\|_\infty \leq x) \geq 1 - 1/n\}, \quad n \geq 1. \tag{2.1}$$

Due to its importance in multivariate extreme value theory, we emphasise here the case of random vectors with non negative components and regular variations on $[0,\infty)^d$. In this simple case, regular variation can be characterised by the convergence of the tail function, see Hult and Lindskog [107]: we have the equivalence

i) the random vector $\mathbf{X}$ is regularly varying on $[0,\infty)^d$ with limit measure Λ, that is

$$n\mathbb{P}(a_n^{-1}\mathbf{X} \in \cdot) \xrightarrow{M_0} \Lambda(\cdot), \qquad \text{as } n \to \infty;$$

i') the tail function $1 - F$ is regularly varying with limit function $V(\mathbf{x}) = \Lambda([\mathbf{0},\mathbf{x}]^c)$, i.e.,

$$\lim_{u \to +\infty} \frac{1 - F(u\mathbf{x})}{1 - F(u\mathbf{1_d})} = V(\mathbf{x}), \quad \mathbf{x} \in [0,\infty)^d \setminus \{\mathbf{0}\};$$

Paralleling the univariate extreme value theory and the equivalence i)-iv) mentioned in the introduction, we consider a sequence $\mathbf{X}, \mathbf{X_1}, \mathbf{X_2}, \ldots$ of non negative random vectors with cumulative distribution F on $[0,\infty)^d$ and we assume for convenience $\mathbb{P}(\mathbf{X} = \mathbf{0}) = 0$. The following statements are known to be equivalent, see e.g. the monograph by Resnick [140] or Coles [57]:

i) (regular variation) the random vector $\mathbf{X}$ is regularly varying on $[0,\infty)^d$ with α-homogeneous limit measure Λ;

ii) (componentwise maxima) the rescaled componentwise maximum $a_n^{-1}\max(\mathbf{X_1},\ldots,\mathbf{X_n})$ converges in distribution as $n \to \infty$ to a jointly α-Fréchet random vector with exponent function $V(\mathbf{x}) = \Lambda([\mathbf{0},\mathbf{x}]^c)$, i.e.,

$$\lim_{n\to\infty} \mathbb{P}\left[a_n^{-1}\max(\mathbf{X_1},\ldots,\mathbf{X_n}) \leq \mathbf{x}\right] = \exp(-V(\mathbf{x})), \quad \mathbf{x} \in [0,\infty)^d \setminus \{\mathbf{0}\};$$

iii) (excess above threshold) the rescaled exceedance $u^{-1}\mathbf{X}$ given that some component of $\mathbf{X}$ exceeds $u > 0$ converges in distribution as $u \to \infty$ to an α-Pareto random vector, i.e.,

$$\lim_{u\to\infty} \mathbb{P}\left[u^{-1}\mathbf{X} \not\leq \mathbf{x} \mid \mathbf{X} \not\leq u\mathbf{1_d}\right] = \frac{V(\mathbf{x} \vee \mathbf{1_d})}{V(\mathbf{1_d})}, \quad \mathbf{x} \in [0,\infty)^d \setminus [0,1]^d;$$

iv) (sample point process) the sample point process $\{a_n^{-1}\mathbf{X_i},\ 1 \leq i \leq n\}$ converges in distribution to a Poisson point process on $[0,\infty)^d \setminus \{\mathbf{0}\}$ with intensity Λ.

2.2.2 A multivariate version of Breiman Lemma

Before considering its multivariate extension, let us recall the celebrated Breiman Lemma (see Breiman [29, Proposition 3]).

Lemma 2.1 (Breiman's lemma)**.** *Let R and Z be independent non negative random variables satisfying either of the following conditions:*

i) the tail function $1 - F$ of R is regularly varying at infinity with index $-\alpha < 0$ and $\mathbb{E}[Z^{\alpha+\varepsilon}] < \infty$ for some $\varepsilon > 0$;

ii) $1 - F(x) \sim Cx^{-\alpha}$ as $x \to \infty$ for some $C > 0$ and $\mathbb{E}[Z^\alpha] < \infty$.

Then, the product RZ is regularly varying with index α and

$$\mathbb{P}(RZ > x) \sim \mathbb{E}[Z^\alpha]\mathbb{P}(R > x) \quad \text{as } x \to \infty.$$

The following multivariate extension of Breiman's Lemma follows the line of Davis and Mikosch [61, section 4.1] .

Proposition 2.1. *Let R be a non negative random variable and $\mathbf{Z}$ an independent d-dimensional random vector. Assume either of the following conditions is satisfied:*

i) the tail function $1 - F$ of R is regularly varying at infinity with index $-\alpha < 0$ and $\mathbb{E}[\|\mathbf{Z}\|^{\alpha+\varepsilon}] < \infty$ for some $\varepsilon > 0$;

ii) $1 - F(x) \sim Cx^{-\alpha}$ as $x \to \infty$ for some $C > 0$ and $\mathbb{E}[\|\mathbf{Z}\|^\alpha] < \infty$.

Then the product $\mathbf{X} = R\mathbf{Z}$ defines a regularly varying random vector on $[-\infty,\infty]^d \setminus \{\mathbf{0}\}$ with index α. More precisely, :

$$n\mathbb{P}(a_n^{-1}\mathbf{X} \in \cdot) \xrightarrow{M_0} \Lambda(\cdot), \qquad \text{in } M_0(\mathbb{R}^d) \text{ as } n \to \infty,$$

where a_n is the quantile of order $1 - 1/n$ of R and the limit measure Λ is homogeneous of order α and given by

$$\Lambda(A) = \int_0^\infty \mathbb{P}(u\mathbf{Z} \in A)\alpha u^{-\alpha-1}\mathrm{d}u, \quad A \subset \mathbb{R}^d \setminus \{\mathbf{0}\} \text{ Borel.} \tag{2.2}$$

Moreover, in the case when $\mathbf{Z}$ is non-negative, Λ is supported by $[0,\infty)^d \setminus \{\mathbf{0}\}$ and we have

$$V(\mathbf{x}) := \Lambda([\mathbf{0},\mathbf{x}]^c) = \mathbb{E}\left[\bigvee_{i=1}^d \left(\frac{Z_i}{x_i}\right)^\alpha\right], \quad \mathbf{x} \in [0,+\infty)^d \setminus \{\mathbf{0}\}.$$

Example 2.1. For example, this applies directly to the multivariate Student distribution with ν degrees of freedom that is the product of an inverse $\chi^2(\nu)$ distribution (with heavy tail of order $\nu/2$) and an independent multivariate Gaussian distributions (with moments of all orders). See Nikololoupolos et al. [133] and Section 2.2.4 below.

Proof of Proposition 2.1. Consider an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^{d-1}$ and denote by $\mathcal{S}^{d-1}$ the unit sphere. For $x > 0$ and $B \subset \mathcal{S}^{d-1}$ Borel, define

$$A = \left\{ z \in \mathbb{R}^d \ : \ \|z\| > x, \ z/\|z\| \in B \right\}. \tag{2.3}$$

We have, as $n \to \infty$,

$$\begin{aligned} n\mathbb{P}(a_n^{-1}\mathbf{X} \in A) &= n\mathbb{P}\left(R\|\mathbf{Z}\| > a_n x, \ \mathbf{Z}/\|\mathbf{Z}\| \in B\right) = n\mathbb{P}\left(R\|\mathbf{Z}\|1_{\{\mathbf{Z}/\|\mathbf{Z}\|\in B\}} > a_n x\right) \\ &\sim n\mathbb{E}\left(\|\mathbf{Z}\|^\alpha 1_{\{\mathbf{Z}/\|\mathbf{Z}\|\in B\}}\right)\mathbb{P}(R > a_n x) \sim \mathbb{E}\left(\|\mathbf{Z}\|^\alpha 1_{\{\mathbf{Z}/\|\mathbf{Z}\|\in B\}}\right) x^{-\alpha} n\mathbb{P}(R > a_n) \\ &\to x^{-\alpha}\mathbb{E}\left(\|\mathbf{Z}\|^\alpha 1_{\{\mathbf{Z}/\|\mathbf{Z}\|\in B\}}\right). \end{aligned} \tag{2.4}$$

We have used here the univariate Breiman Lemma 2.1 to go from the first to the second line and then the fact that R has a regularly varying tail with index $\alpha > 0$ and that $n\mathbb{P}(R > a_n) \to 1$. Using the fact that the sets of the form (2.3) form a convergence determining class (Hult and Linskog [107]), we deduce from Equation (2.4) the M_0-convergence $n\mathbb{P}(\mathbf{X}/a_n \in \cdot) \xrightarrow{M_0} \Lambda(\cdot)$, where the limit measure Λ is characterised by

$$\Lambda(A) = x^{-\alpha}\mathbb{E}\left(\|\mathbf{Z}\|^\alpha 1_{\{\mathbf{Z}/\|\mathbf{Z}\|\in B\}}\right) \tag{2.5}$$

for all set A of the form (2.3). We then check that Λ admits the integral representation (2.2). Computing the right hand side of (2.2) with A given by (2.3), we get

$$\begin{aligned} \int_0^\infty \mathbb{P}(u\mathbf{Z} \in A)\alpha u^{-\alpha-1}\mathrm{d}u &= \int_0^\infty \mathbb{E}\left(1_{\{u\|\mathbf{Z}\|>x, \mathbf{Z}/\|\mathbf{Z}\|\in B\}}\right)\alpha u^{-\alpha-1}\mathrm{d}u \\ &= \mathbb{E}\left(1_{\{\mathbf{Z}/\|\mathbf{Z}\|\in B\}}\int_0^\infty 1_{\{u>x/\|\mathbf{Z}\|\}}\alpha u^{-\alpha-1}\mathrm{d}u\right) \\ &= x^{-\alpha}\mathbb{E}\left(\|\mathbf{Z}\|^\alpha 1_{\{\mathbf{Z}/\|\mathbf{Z}\|\in B\}}\right) = \Lambda(A). \end{aligned}$$

Since the sets A of the form (2.3) form a determining class, the integral representation (2.2) holds for all $A \subset \mathbb{R}^d \setminus \{\mathbf{0}\}$ Borel. We can then check directly that Λ is homogeneous: for $v > 0$,

$$\begin{aligned} \Lambda(vA) &= \int_0^\infty \mathbb{P}(u\mathbf{Z} \in vA)\alpha u^{-\alpha-1}\mathrm{d}u = \int_0^\infty \mathbb{P}(v^{-1}u\mathbf{Z} \in A)\alpha u^{-\alpha-1}\mathrm{d}u \\ &= v^{-\alpha}\int_0^\infty \mathbb{P}(u\mathbf{Z} \in A)\alpha u^{-\alpha-1}\mathrm{d}u = v^{-\alpha}\Lambda(A), \end{aligned}$$

where we used the change of variable $u' = u/v$ on the second line.

Finally, when $\mathbf{Z}$ is supported by $[0,\infty)^d$, Equation (2.2) implies that Λ is supported by $[0,\infty)^d \setminus \{\mathbf{0}\}$ and the tail function V is computed as follows:

$$\begin{aligned}
V(\mathbf{x}) := \Lambda([\mathbf{0},\mathbf{x}]^c) &= \int_{[0,\infty)^d} \mathbb{P}(u\mathbf{Z} \notin [\mathbf{0},\mathbf{x}])\alpha u^{-\alpha-1}\mathrm{d}u \\
&= \int_{[0,\infty)^d} \mathbb{P}\left(u > Z_i x_i \text{ for some } 1 \le i \le d\right)\alpha u^{-\alpha-1}\mathrm{d}u \\
&= \int_{[0,\infty)^d} \mathbb{P}\left(u > \min_{1\le i\le d} \frac{x_i}{Z_i}\right)\alpha u^{-\alpha-1}\mathrm{d}u = \mathbb{E}\left[\left(\min_{1\le i\le d} \frac{x_i}{Z_i}\right)^{-\alpha}\right] \\
&= \mathbb{E}\left[\bigvee_{i=1}^{d}\left(\frac{Z_i}{x_i}\right)^{\alpha}\right].
\end{aligned}$$

□

Proposition 2.2. *If $\mathbf{Z}$ has a density $f_{\mathbf{Z}}$, then Λ is absolutely continuous with respect to the Lebesgue measure and its Radon-Nikodym derivative is given by*

$$\lambda(\mathbf{z}) = \int_0^\infty f_Z\left(\mathbf{z}/u\right)\alpha u^{-d-\alpha-1}\mathrm{d}u. \tag{2.6}$$

and is homogeneous of order $-d-\alpha$, that is

$$\lambda(v\mathbf{z}) = v^{-d-\alpha}\lambda(\mathbf{z}), \quad v > 0,\ \mathbf{z} \in \mathbb{R}^d \setminus \{\mathbf{0}\}. \tag{2.7}$$

Proof. If $\mathbf{Z}$ has a density $f_{\mathbf{Z}}$, the measure Λ writes

$$\begin{aligned}
\Lambda(A) &= \int_0^\infty \mathbb{P}(u\mathbf{Z} \in A)\alpha u^{-\alpha-1}\mathrm{d}u = \int_0^\infty \int_{\mathbb{R}^d} 1_{\{u\mathbf{z}\in A\}} f_{\mathbf{Z}}(\mathbf{z})\mathrm{d}\mathbf{z}u^{-\alpha-1}\mathrm{d}u \\
&= \int_0^\infty \int_A f_{\mathbf{Z}}\left(\mathbf{z}/u\right)\alpha u^{-\alpha-d-1}\mathrm{d}\mathbf{z}\mathrm{d}u = \int_A \lambda(\mathbf{z})\mathrm{d}\mathbf{z},
\end{aligned}$$

where we use the change of variable $\mathbf{z}' = u\mathbf{z}$ and Fubini Theorem. Furthermore, with the change of variable $u' = u/v$,

$$\lambda(v\mathbf{z}) = \int_0^\infty f_{\mathbf{Z}}(v\mathbf{z}/u)\alpha u^{-d-\alpha-1}\mathrm{d}u = v^{-d-\alpha}\int_0^\infty f_{\mathbf{Z}}(\mathbf{z}/u)\alpha u^{-d-\alpha-1}\mathrm{d}u = v^{-d-\alpha}\lambda(\mathbf{z}).$$

□

2.2.3 A copula point of view

When focusing on the dependence structure, Proposition 2.1 can be rephrased in terms of copulas (we refer to Joe [114] for a background on copulas and Gudendorf and Segers for

extreme value copulas [94]). Following Krupskii et al. [117], we consider here the simple common factor model

$$\mathbf{X} = \alpha E \mathbf{1_d} + \mathbf{Y} \tag{2.8}$$

with $\alpha > 0$, E exponentially distributed and, independently, $\mathbf{Y}$ a d-dimensional random vector such that $\mathbb{E}[e^{\alpha Y_i}] < \infty$, for all $i = 1, \dots, d$. The different component of $\mathbf{X}$ share the common factor E that introduces dependence in the extremes, because the components of $\mathbf{Y}$ are lighter tailed. Since the exponential distribution has a density, all the components $X_i = \alpha E + Y_i$ have a continuous distribution. Sklar Theorem entails that the copula $C_{\mathbf{X}}$ pertaining to $\mathbf{X}$ is uniquely defined by

$$C_{\mathbf{X}}(u_1, \dots, u_d) = F_{\mathbf{X}}(F^{\leftarrow}_{X_i}(u_1), \dots, F^{\leftarrow}_{X_d}(u_d)), \quad (u_1, \dots, u_d) \in [0,1]^d,$$

where $F_{\mathbf{X}}$ denotes the multivariate cumulative distribution of $\mathbf{X}$ and $F^{\leftarrow}_{X_i}$ the quantile function of component X_i.

Proposition 2.3. *Consider the copula $C_{\mathbf{X}}$ associated to the random vector $\mathbf{X}$ defined by* (2.8)*. Then*

$$C^n_{\mathbf{X}}(u_1^{1/n}, \dots, u_d^{1/n}) \to C_{\mathbf{V}}(u_1, \dots, u_d), \quad (u_1, \dots, u_d) \in [0,1]^d,$$

where

$$C_{\mathbf{V}}(u_1, \dots, u_d) = \exp\left(-V(\sigma_1(-\log u_1)^{1/\alpha}, \dots, \sigma_d(-\log u_d)^{1/\alpha})\right)$$

with

$$\sigma_i^\alpha = \mathbb{E}[e^{\alpha Y_i}] \quad \textit{and} \quad V(\mathbf{x}) = \mathbb{E}\left[\vee_{i=1}^d \frac{e^{\alpha Y_i}}{x_i^\alpha}\right].$$

In words, $C_{\mathbf{X}}$ belongs to the domain of attraction of the extreme value copula $C_{\mathbf{V}}$.

Here, we use the fact that $\exp(\alpha E)$ has an α-Pareto distribution but, in view of the proof and the multivariate Breiman Lemma, the result holds as soon as $\exp(\alpha E)$ has an heavy tail with index α and $(\alpha + \varepsilon)Y_i$ has a finite exponential moment for $i = 1, \dots, d$.

Proof of Proposition 2.3. By Proposition 2.1, $e^{\mathbf{X}} = e^{\alpha E} e^{\mathbf{Y}}$ is regularly varying with exponent function V and hence, the normalised maximum of n independent copies of $\mathbf{X}$ converge to an α-Fréchet vector with distribution function $e^{-V(\mathbf{x})}$. On the other hand, since the exponential transformation operates separately on each component, $e^{\mathbf{X}}$ has copula $C_{\mathbf{X}}$ and the normalised maximum of n i.i.d. copies has copula $C^n_{\mathbf{X}}(u_1^{1/n}, \dots, u_d^{1/n})$. It remains to note that $C_{\mathbf{V}}$ is the copula associated with the limiting α-Fréchet vector, where the i-th margin as shape parameter α and scale parameter σ. The fact that convergence of point-wise maxima implies convergence of the copula is justified in Deheuvels [63]. □

2.2.4 Examples

In this section, we apply Proposition 2.1 and consider various models for the various random vector $\mathbf{Z}$. For these models, we provide an explicit expression for the limit measure Λ that characterises the regular variation of the product $\mathbf{X} = R\mathbf{Z}$. Our computations rely on the general form of the density λ expressed in Proposition 2.2 and technical computations. In the following examples, R is regularly varying with index α.

Gaussian case

The following result states a regular variation result in connection with the extremal-t model, see Nikoloulopoulas et al. [133].

Proposition 2.4. *In the framework of the multivariate Breiman's lemma, if* $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$*, then the limit measure* Λ *has density*

$$\lambda(\mathbf{z}) = \frac{\alpha}{(2\pi)^{d/2}|\Sigma|^{1/2}} \Gamma\left(\frac{\alpha+d}{2}\right) \left(\frac{\mathbf{z}^\top \Sigma^{-1} \mathbf{z}}{2}\right)^{-(\alpha+d)/2}, \quad \mathbf{z} \in \mathbb{R}^d \setminus \{\mathbf{0}\}.$$

Proof. Starting from Eq. (2.6) and introducing the Gaussian density, we get

$$\begin{aligned}
\lambda(\mathbf{z}) &= \int_0^\infty f_{\mathbf{Z}}\left(\frac{\mathbf{z}}{u}\right) \alpha u^{-\alpha-d-1} \mathrm{d}u \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}^d |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2u^2} \mathbf{z}^\top \Sigma^{-1} \mathbf{z}\right\} \alpha u^{-\alpha-d-1} \mathrm{d}u.
\end{aligned}$$

The change of variable $v = 1/u$ in the integral yields

$$\begin{aligned}
\lambda(\mathbf{z}) &= (2\pi)^{-d/2} |\Sigma|^{-1/2} \alpha \int_0^\infty \exp\left\{-\frac{v^2}{2} \mathbf{z}^\top \Sigma^{-1} \mathbf{z}\right\} u^{\alpha+d-1} \mathrm{d}u \\
&= (2\pi)^{-d/2} |\Sigma|^{-1/2} \alpha \frac{2}{\mathbf{z}^\top \Sigma^{-1} \mathbf{z}} \int_0^\infty \frac{\mathbf{z}^\top \Sigma \mathbf{z}}{2} \exp\left\{-\frac{v^2}{2} \mathbf{z}^\top \Sigma^{-1} \mathbf{z}\right\} u^{\alpha+d-1} \mathrm{d}u \\
&= (2\pi)^{-d/2} |\Sigma|^{-1/2} \alpha \frac{2}{\mathbf{z}^\top \Sigma^{-1} \mathbf{z}} \mathbb{E}\left[X^{\alpha+d-2}\right]
\end{aligned}$$

where X has a Weibull distribution with shape parameter equal to 2 and scale parameter equal to $\sqrt{2/(z^\top \Sigma^{-1} z)}$. We deduce

$$\mathbb{E}\left[X^{\alpha+d-2}\right] = \left(\frac{\mathbf{z}^\top \Sigma^{-1} \mathbf{z}}{2}\right)^{-(\alpha+d-2)/2} \Gamma\left(\frac{\alpha+d}{2}\right)$$

and we obtain the claimed formula for $\lambda(\mathbf{z})$. □

Log-normal case

The case of log-normal spectral functions is connected with the Hüsler-Reiss model [110], see also Wadsworth and Tawn [179].

Proposition 2.5. *In the framework of the multivariate Breiman's lemma, if* $\mathbf{Z} \sim \mathcal{LN}(\mathbf{m}, \Sigma)$ *with* Σ *definite positive, then the limit measure* Λ *has density*

$$\lambda(\mathbf{z}) = C \exp\left\{-\frac{1}{2} \log \mathbf{z}^\top Q \log \mathbf{z} + \boldsymbol{l} \log \mathbf{z}\right\} \prod_{i=1}^d z_i^{-1}, \quad \mathbf{z} \in (0, \infty)^d,$$

where

$$C = \frac{\alpha}{(2\pi)^{(d-1)/2}|\Sigma|^{1/2}\sqrt{\mathbf{1_d}^\top\Sigma^{-1}\mathbf{1_d}}}\exp\left\{-\frac{1}{2}\mathbf{m}^\top\Sigma^{-1}\mathbf{m}+\frac{1}{2}\frac{(\mathbf{m}^\top\Sigma^{-1}\mathbf{1_d}-\alpha)^2}{\mathbf{1_d}^\top\Sigma^{-1}\mathbf{1_d}}\right\},$$

$$Q = \Sigma^{-1} - \frac{\Sigma^{-1}\mathbf{1_d}\mathbf{1_d}^\top\Sigma^{-1}}{\mathbf{1_d}^\top\Sigma^{-1}\mathbf{1_d}}, \tag{2.9}$$

$$\boldsymbol{l} = \left(\mathbf{m}^\top - \frac{\alpha+\mathbf{m}^\top\Sigma^{-1}\mathbf{1_d}}{\mathbf{1_d}^\top\Sigma^{-1}\mathbf{1_d}}\mathbf{1_d}^\top\right)\Sigma^{-1}. \tag{2.10}$$

and

$$V(\mathbf{x}) = \frac{C(2\pi)^{(d-1)/2}}{\alpha}\sum_{i=1}^{d} x_i^{-\alpha}|Q_{-i}|^{-1/2}\exp\left\{\frac{1}{2}\boldsymbol{l}_{-i}^\top Q_{-i}^{-1}\boldsymbol{l}_{-i}\right\}\Phi_{d-1}\left(\log\frac{x_{-i}}{x_i};Q_{-i}^{-1}\boldsymbol{l}_{-i},Q_{-i}^{-1}\right).$$

Proof. Starting from Eq. (2.6) and introducing the log-normal Gaussian density, we get

$$\begin{aligned}\lambda(\mathbf{z}) &= \int_0^\infty f_{\mathbf{Z}}\left(\frac{\mathbf{z}}{u}\right)\alpha u^{-\alpha-d-1}\mathrm{d}u\\ &= \int_0^\infty \prod_{i=1}^{d} z_i^{-1}\alpha|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(\log(\mathbf{z})-\log(u)\mathbf{1_d}-\mathbf{m})^\top\Sigma^{-1}(\log(\mathbf{z})-\log(u)\mathbf{1_d}-\mathbf{m})\right\}\\ &\quad (2\pi)^{-d/2}u^{-\alpha-1}\mathrm{d}u\end{aligned}$$

The change of variable $v=\log(u)$ yields

$$\lambda(\mathbf{z}) = \alpha|\Sigma|^{-1/2}(2\pi)^{-d/2}\prod_{i=1}^{d} z_i^{-1}\int_{-\infty}^{\infty}\exp\left\{\mathrm{P}(v)\right\}\mathrm{d}v$$

with

$$\begin{aligned}\mathrm{P}(v) &= -\frac{1}{2}(\log(\mathbf{z})-v\mathbf{1_d}-\mathbf{m})^\top\Sigma^{-1}(\log(\mathbf{z})-v\mathbf{1_d}-\mathbf{m})-\alpha v\\ &= -\frac{1}{2}\mathbf{1_d}^\top\Sigma^{-1}\mathbf{1}_d v^2+\left(\log(\mathbf{z})^\top\Sigma^{-1}\mathbf{1_d}-\mathbf{m}^\top\Sigma^{-1}\mathbf{1_d}-\alpha\right)v\\ &\quad -\frac{1}{2}\log(\mathbf{z})^\top\Sigma^{-1}\log(\mathbf{z})-\frac{1}{2}\mathbf{m}^\top\Sigma^{-1}\mathbf{m}+\log(\mathbf{z})^\top\Sigma^{-1}\mathbf{m}\\ &= -\frac{1}{2}C_1v^2+C_2v+C_3.\end{aligned}$$

Recognising a Gaussian integral, we get with $X\sim\mathcal{N}(0,C_1^{-1})$,

$$\int_{-\infty}^{\infty}\exp\left\{P(v)\right\}\mathrm{d}v = \sqrt{\frac{2\pi}{C_1}}e^{C_3}\mathbb{E}[\exp\{C_2X\}] = \sqrt{\frac{2\pi}{C_1}}e^{C_3}e^{\frac{C_2^2}{2C_1}}.$$

We deduce the claimed formula for $\lambda(\mathbf{z})$ after some straightforward simplifications. □

Independent Fréchet case

The case of independent spectral components is related to the logistic model [94].

Proposition 2.6 (Frechet case)**.** *Suppose* $\mathbf{Z} = (Z_1, \ldots, Z_d)$ *with* $Z_i \sim \text{Frechet}(\lambda_i, \beta)$ *independent with* $\beta > \alpha$*. Then, the limit measure* Λ *in multivariate Breiman's lemma has density*

$$\lambda(\mathbf{z}) = \alpha\beta^{d-1}\Gamma(d-\alpha/\beta)\prod_{i=1}^{d}\frac{z_i^{-\beta-1}}{\lambda_i^{-\beta}}\left(\sum_{i=1}^{d}\left(\frac{z_i}{\lambda_i}\right)^{-\beta}\right)^{(\alpha+1)/\beta-d}$$

with Γ *the Gamma function and*

$$V(\mathbf{x}) := \Lambda([\mathbf{0},\mathbf{x}]^c) = \Gamma\left(1-\frac{\alpha}{\beta}\right)\left(\sum_{i=1}^{d}\left(\frac{x_i}{\lambda_i}\right)^{-\beta}\right)^{\alpha/\beta}.$$

Proof. Starting from Eq. (2.6) and introducing the product Fréchet density yields

$$\begin{aligned}
\lambda(\mathbf{z}) &= \int_0^\infty f_{\mathbf{Z}}\left(\frac{\mathbf{z}}{u}\right)\alpha u^{-\alpha-d-1}\mathrm{d}u \\
&= \int_0^\infty \prod_{i=1}^{d}\left(z_i^{-\beta-1}u^{\beta+1}\beta\lambda_i^{\beta}\exp\left\{-(z_i/\lambda_i u)^{-\beta}\right\}\right)\alpha u^{-\alpha-d-1}\mathrm{d}u \\
&= \alpha\beta^d\prod_{i=1}^{d}\frac{z_i^{-\beta-1}}{\lambda_i^{-\beta}}\int_0^\infty u^{-\alpha+\beta d-1}\exp\left\{-u^\beta\sum_{i=1}^{d}\frac{z_i}{\lambda_i}^{-\beta}\right\}\mathrm{d}u.
\end{aligned}$$

The change of variable $v = u^\beta\sum_{i=1}^{d}\left(\frac{z_i}{\lambda_i}\right)^{-\beta}$ in the integral gives

$$\lambda(\mathbf{z}) = \alpha\beta^d\prod_{i=1}^{d}\frac{z_i^{-\beta-1}}{\lambda_i^{-\beta}}\frac{1}{\beta}\left(\sum_{i=1}^{d}\left(\frac{z_i}{\lambda_i}\right)^{-\beta}\right)^{(\alpha+1)/\beta-d}\int_0^\infty e^{-v}v^{d-\alpha/\beta-1}\mathrm{d}v$$

The last integral is the definition of the Gamma function $\Gamma(d-\alpha/\beta)$. Proposition 2.1 gives

$$\begin{aligned}
V(\mathbf{x}) &= \mathbb{E}\left[\bigvee_{i=1}^{d}\left(\frac{Z_i}{x_i}\right)^{\alpha}\right] \\
&= \int_0^\infty \mathbb{P}\left(\bigvee_{i=1}^{d}\left(\frac{Z_i}{x_i}\right)^{\alpha} > x\right)\mathrm{d}x \\
&= \int_0^\infty \left[1-\prod_{i=1}^{d}\mathbb{P}\left(\left(\frac{Z_i}{x_i}\right)^{\alpha}\leq x\right)\right]\mathrm{d}x.
\end{aligned}$$

Introducing the Fréchet density function yields

$$V(\mathbf{x}) = \int_0^\infty \left[1 - \exp\left(-x^{-\beta/\alpha} \sum_{i=1}^d \left(\frac{x_i}{\lambda_i}\right)^{-\beta}\right)\right] \mathrm{d}x.$$

The change of variable $y = x\left(\sum_{i=1}^d \left(\frac{x_i}{\lambda_i}\right)^{-\beta}\right)^{-\alpha/\beta}$ gives

$$V(\mathbf{x}) = \left(\sum_{i=1}^d \left(\frac{x_i}{\lambda_i}\right)^{-\beta}\right)^{\alpha/\beta} \int_0^\infty \left[1 - \exp\left(-y^{-\beta/\alpha}\right)\right] \mathrm{d}y$$

The last integral corresponds to the expectation of a Fréchet$(1, \beta/\alpha)$ and therefore, assuming $\beta > \alpha$, we have the result. □

Independent Weibull case

The case of independent spectral components is related to the negative logistic model [94].

Proposition 2.7 (Weibull case)**.** *Suppose* $\mathbf{Z} = (Z_1, \ldots, Z_d)$ *with* $Z_i \sim$ Weibull(λ_i, β) *independent with* $\alpha > \beta$*. Then the limit measure* Λ *in multivariate Breiman's Lemma has density*

$$\lambda(\mathbf{z}) = \alpha\beta^{d-1}\Gamma(d + \alpha/\beta) \left(\sum_{i=1}^d \left(\frac{z_i}{\lambda_i}\right)^\beta\right)^{-(\alpha+1)/\beta - d} \prod_{i=1}^d \frac{z_i^{\beta-1}}{\lambda_i^\beta}$$

and

$$V(\mathbf{x}) := \Lambda([\mathbf{0}, \mathbf{x}]^c) = \Gamma\left(1 + \frac{\alpha}{\beta}\right) \sum_{\emptyset \neq J \subset \{1, \cdots, d\}} (-1)^{|J|+1} \left(\sum_{j \in J} \left(\frac{x_j}{\lambda_j}\right)^\beta\right)^{-\alpha/\beta}.$$

Proof. Starting from Eq. (2.6) and introducing the product Weibull density yields

$$\begin{aligned}
\lambda(\mathbf{z}) &= \int_0^\infty f_{\mathbf{Z}}\left(\frac{\mathbf{z}}{u}\right) \alpha u^{-\alpha-d-1} \mathrm{d}u \\
&= \int_0^\infty \prod_{i=1}^d \left(\frac{\beta}{\lambda_i}\left(\frac{z_i}{u\lambda_i}\right)^{\beta-1} \exp\left\{-\left(\frac{z_i}{u\lambda_i}\right)^\beta\right\}\right) \alpha u^{-\alpha-d-1} \mathrm{d}u \\
&= \alpha\beta^d \prod_{i=1}^d \left(\frac{1}{\lambda_i}\left(\frac{z_i}{\lambda_i}\right)^{\beta-1}\right) \int_0^\infty \exp\left\{-u^{-\beta}\left(\sum_{i=1}^d \left(\frac{z_i}{\lambda_i}\right)^\beta\right)\right\} u^{-\alpha-\beta d-1} \mathrm{d}u.
\end{aligned}$$

The change of variable $v = u^{-\beta}\left(\sum_{i=1}^d \left(\frac{z_i}{\lambda_i}\right)^\beta\right)$ in the integral gives

$$\lambda(\mathbf{z}) = \alpha\beta^{d-1} \left(\sum_{i=1}^d \left(\frac{z_i}{\lambda_i}\right)^\beta\right)^{-(\alpha+1)/\beta - d} \prod_{i=1}^d \frac{z_i^{\beta-1}}{\lambda_i^\beta} \int_0^\infty e^{-v} v^{\frac{\alpha}{\beta}+d-1} \mathrm{d}v.$$

Proposition 2.1 yields

$$\begin{aligned} V(\mathbf{x}) &= \mathbb{E}\left[\bigvee_{i=1}^{d}\left(\frac{Z_i}{x_i}\right)^{\alpha}\right] \\ &= \int_0^\infty \left[1-\prod_{i=1}^{d}\mathbb{P}\left(\left(\frac{Z_i}{x_i}\right)^{\alpha}\leq x\right)\right]\mathrm{d}x. \end{aligned}$$

Introducing the Weibull density function yields

$$\begin{aligned} V(\mathbf{x}) &= \int_0^\infty \left[1-\prod_{i=1}^{d}\left(1-\exp\left(-x^{\beta/\alpha}\left(\frac{x_i}{\lambda_i}\right)^{\beta}\right)\right)\right]\mathrm{d}x \\ &= \int_0^\infty \sum_{\emptyset\neq J\subset\{1,\cdots,d\}}(-1)^{|J|+1}\exp\left(-x^{\beta/\alpha}\sum_{j\in J}\left(\frac{x_j}{\lambda_j}\right)^{\beta}\right)\mathrm{d}x. \end{aligned}$$

The change of variable $y = x\left(\sum_{j\in J}\left(\frac{x_j}{\lambda_j}\right)^{\beta}\right)^{\alpha/\beta}$ yields

$$V(\mathbf{x}) = \sum_{\emptyset\neq J\subset\{1,\cdots,d\}}(-1)^{|J|+1}\left(\sum_{j\in J}\left(\frac{x_j}{\lambda_j}\right)^{\beta}\right)^{-\alpha/\beta}\int_0^\infty \exp\left(-y^{\beta/\alpha}\right)\mathrm{d}y.$$

The last integral correspond to the expectation of a Weibull$(1,\beta/\alpha)$. □

Independent Gamma case

This last example is related to the max-stable model with Dirichlet spectral density in the case where $\beta_i \equiv 1$. The restriction of λ on the simplex is proportional to the Dirichlet density.

Proposition 2.8 (Gamma case)**.** *Suppose* $\mathbf{Z} = (Z_1,\ldots,Z_d)$ *with* $Z_i \sim \Gamma(\theta_i,\beta_i)$ *independent. Then*

$$\lambda(\mathbf{z}) = \alpha\Gamma\left(\alpha+\sum_{i=1}^{d}\theta_i\right)\left(\sum_{i=1}^{d}\beta_i z_i\right)^{-\sum_{i=1}^{d}\theta_i-\alpha}\prod_{i=1}^{d}\left(\frac{\beta_i^{\theta_i}z_i^{\theta_i-1}}{\Gamma(\theta_i)}\right)$$

Proof.

$$\begin{aligned} \lambda(\mathbf{z}) &= \int_0^\infty \prod_{i=1}^{d}\left(\frac{\beta_i^{\theta_i}}{\Gamma(\theta_i)}\left(\frac{z_i}{u}\right)^{\theta_i-1}e^{-\beta_i z_i/u}\right)\alpha u^{-\alpha-1-d}\mathrm{d}u \\ &= \alpha\prod_{i=1}^{d}\left(\frac{\beta_i^{\theta_i}z_i^{\theta_i-1}}{\Gamma(\theta_i)}\right)\int_0^\infty u^{-\sum_{i=1}^{d}\theta_i-\alpha-1}\exp\left\{-u^{-1}\sum_{i=1}^{d}\beta_i z_i\right\}\mathrm{d}u. \end{aligned}$$

Setting $v = u^{-1}\sum_{i=1}^d \beta_i z_i$, we obtain

$$\lambda(\mathbf{z}) = \alpha\left(\sum_{i=1}^d \beta_i z_i\right)^{-\sum_{i=1}^d \theta_i - \alpha} \prod_{i=1}^d \left(\frac{\beta_i^{\theta_i} z_i^{\theta_i - 1}}{\Gamma(\theta_i)}\right) \int_0^\infty e^{-v} v^{\sum_{i=1}^d \theta_i + \alpha - 1} \mathrm{d}u.$$

□

2.2.5 Non standard regular variation

Following Resnick [141], non-standard multivariate regular variations correspond to different tail index for the different components. Proposition 2.1 has a simple extension to this case.

Proposition 2.9. *Let R be a non negative heavy-tailed random variable with index 1, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_d) \in (0,\infty)^d$ and $\mathbf{Z} = (Z_1, \cdots, Z_d)$ a d-dimensional random vector such that $\mathbb{E}|Z_i|^{\alpha_i + \varepsilon} < \infty$ for some $\varepsilon > 0$. Then the product $\mathbf{X} = (R^{1/\alpha_1} Z_1, \cdots, R^{1/\alpha_d} Z_d) = R^{1/\alpha}\mathbf{Z}$ satisfies*

$$n\mathbb{P}(a_n^{-1/\alpha}\mathbf{X} \in \cdot) \xrightarrow{M_0} \Lambda(\cdot)$$

where a_n is the quantile of order $1 - n^{-1}$ of R and the limit measure Λ satisfies

$$\Lambda(A) = \int_0^\infty \mathbb{P}\left(u^{-1/\alpha}\mathbf{Z} \in A\right) \mathrm{d}u, \quad A \subset \mathbb{R}^d \setminus \{\mathbf{0}\} \text{ measurable.} \tag{2.11}$$

Proof. Proposition 2.1 for $\tilde{\mathbf{X}} = R\mathbf{Z}^\alpha = (RZ_1^{\alpha_1}, \ldots, RZ_d^{\alpha_d})$ yields the regular variations for $\tilde{\mathbf{X}}$. Then, the change of variable $\mathbf{X} = \tilde{\mathbf{X}}^{1/\alpha}$ together with the continuous mapping theorem for M_0-convergence [107] imply the non-standard regular variations stated in Proposition 2.9 . □

Chapter 3

On the Hüsler-Reiss Pareto distribution

3.1 Introduction

Following the chapter 2, we propose a thorough study of the so-called Hüsler-Reiss Pareto model, that is the exceedance Pareto model associated with the max-stable Hüsler-Reiss model [110]. The exceedances of the related Brown-Resnick spatial model were considered recently by Wadsworth and Tawn [179] who proposed inference via censored maximum likelihood, see also Kiriliouk et al. [116]. Here, we focus on the finite-dimensional multivariate Hüsler-Reiss Pareto model and notice that it has a simple exponential family structure (see Barndorff-Nielsen [14]), that seems to have been overlooked in the literature. We propose in Section 3.2 an extensive study of this exponential family structure and consider also maximum likelihood inference as well as perfect simulation. We extend these results in Section 3.3 where we consider the non-standard Hüsler-Reiss Pareto model that incorporates different tail parameters for the different margins. Maximum likelihood estimators are shown again to be asymptotically normal and an alternating optimisation procedure is considered. To conclude, we propose a maximum likelihood ratio test for testing the equality of the different marginal tail parameters.

3.2 The Hüsler-Reiss Pareto model

3.2.1 Definition and transformation properties

Motivated by Proposition 2.5, we introduce the family of Hüsler-Reiss Pareto distributions and study their properties. The main reason why we focus on that particular class is that it enjoys an exponential family property, see Barndorff-Nielsen [14].

Definition 3.1. *Let $d \geq 2$, $\mathbf{a} = (a_1, \ldots, a_d) \in (0, \infty)^d$, $Q \in \mathbb{R}^{d \times d}$ a symmetric positive semi-definite matrix such that* $\operatorname{Ker} Q = \operatorname{span}(\mathbf{1_d})$ *and $\boldsymbol{l} \in \mathbb{R}^d$ such that $\boldsymbol{l}^\top \mathbf{1_d} < 0$. The Hüsler-Reiss*

Pareto model on $[0,\infty)^d \setminus [\mathbf{0},\mathbf{a}]$ *with parameters* $(Q,\boldsymbol{l})$ *is defined by the density*

$$f_{\mathbf{a}}(\mathbf{z};Q,\boldsymbol{l}) = \frac{1}{C_{\mathbf{a}}(Q,\boldsymbol{l})} \exp\left(-\frac{1}{2}\log\mathbf{z}^\top Q \log\mathbf{z} + l^\top \log\mathbf{z}\right)\left(\prod_{i=1}^d z_i^{-1}\right) 1_{\{\mathbf{z}\not\leq\mathbf{a}\}}, \quad \mathbf{z}\in(0,\infty)^d,$$

with $C_{\mathbf{a}}(Q,\boldsymbol{l})$ *the normalisation constant. We call* $\alpha = -\boldsymbol{l}^\top\mathbf{1_d} > 0$ *the exponent of the Pareto distribution* $f_{\mathbf{a}}(\mathbf{z};Q,\boldsymbol{l})$.
We write $Z \rightsquigarrow \mathrm{HRPar}_a(Q,\boldsymbol{l})$ *for a random vector* $\mathbf{Z}$ *with density* $f_{\mathbf{a}}(\mathbf{z};Q,\boldsymbol{l})$.

Remark 3.1. The Hüsler-Reiss Pareto model is closely connected with the exponent measure λ obtained in Proposition 2.5. Indeed, the parameters Q and $\boldsymbol{l}$ introduced there satisfy the constraint stated in Definition 3.1. The symmetric semi-definite positive matrix Q satisfies

$$Q\mathbf{1_d} = \left(\Sigma^{-1} - \frac{\Sigma^{-1}\mathbf{1_d}\mathbf{1_d}^\top\Sigma^{-1}}{\mathbf{1_d}^\top\Sigma^{-1}\mathbf{1_d}}\right)\mathbf{1_d} = \mathbf{0}$$

and, for all vector $\mathbf{x}\in\mathbb{R}^d\setminus\{\mathbf{0}\}$ such that $\mathbf{x}^\top\Sigma^{-1}\mathbf{1_d} = 0$, we have $\mathbf{x}^\top Q\mathbf{x} > 0$ whence we deduce $\operatorname{Ker} Q = \operatorname{span}(\mathbf{1_d})$. As for $\boldsymbol{l}$, we check readily

$$\boldsymbol{l}^\top\mathbf{1_d} = \left(\mathbf{m}^\top - \frac{\alpha + \mathbf{m}^\top\Sigma^{-1}\mathbf{1_d}}{\mathbf{1_d}^\top\Sigma^{-1}\mathbf{1_d}}\mathbf{1_d}^\top\right)\Sigma^{-1}\mathbf{1_d} = -\alpha < 0.$$

Conversely, for all $(Q,\boldsymbol{l})$ as in Definition 3.1, there exist (non unique) $\Sigma\in\mathbb{R}^{d\times d}$ and $\mathbf{m}\in\mathbb{R}^d$ such that Equations (2.9) and (2.10) are satisfied.

Example 3.1. In dimension $d=2$, the model parameters are

$$Q = \begin{pmatrix} c & -c \\ -c & c \end{pmatrix} \quad \text{and} \quad \boldsymbol{l} = \begin{pmatrix} l_1 \\ l_2 \end{pmatrix} \quad \text{with } c>0,\ l_1+l_2<0.$$

The exponent is $\alpha = -(l_1+l_2) > 0$ and

$$f_{\mathbf{a}}(\mathbf{z};Q,\boldsymbol{l}) = \frac{1}{C_{\mathbf{a}}(Q,\boldsymbol{l})}\exp\left(-\frac{c}{2}(\log z_1 - \log z_2)^2 + l_1\log z_1 + l_2\log z_2\right)\frac{1}{z_1z_2}1_{\{\mathbf{z}\not\leq\mathbf{a}\}}.$$

Interestingly, Hüsler-Reiss Pareto distributions inherit from log-normal distributions a stability property under scale and power transformations.

Proposition 3.1. *Let* $\mathbf{Z}\rightsquigarrow\mathrm{HRPar}_{\mathbf{a}}(Q,\boldsymbol{l})$.

(i) For all $\mathbf{u}\in(0,\infty)^d$, $\mathbf{uZ}\rightsquigarrow\mathrm{HRPar}_{\mathbf{ua}}(Q,\boldsymbol{l}+Q\log\mathbf{u})$.

(ii) For all $\beta>0$, $\mathbf{Z}^\beta\rightsquigarrow\mathrm{HRPar}_{\mathbf{a}^\beta}(\beta^{-2}Q,\beta^{-1}\boldsymbol{l})$. *In particular, if* $\mathbf{Z}$ *has exponent* α, $\mathbf{Z}^\beta$ *has exponent* α/β.

Proof. The change of variable $\tilde{\mathbf{z}} = \mathbf{u}\mathbf{z}$ implies

$$\mathbb{P}(\mathbf{u}\mathbf{Z} \in A) = \int_A f_{\mathbf{a}}(\tilde{\mathbf{z}}/\mathbf{u}; Q, \boldsymbol{l}) \prod_{i=1}^d u_i^{-1} \mathrm{d}\tilde{\mathbf{z}}.$$

Simple computations show that

$$\begin{aligned}
&f_{\mathbf{a}}(\mathbf{z}/\mathbf{u}; Q, \boldsymbol{l}) \prod_{i=1}^d u_i^{-1} \\
&\quad = \frac{1}{C_{\mathbf{a}}(Q, \boldsymbol{l})} \exp\left(-\frac{1}{2}(\log \mathbf{z} - \log \mathbf{u})^\top Q (\log \mathbf{z} - \log \mathbf{u}) + \boldsymbol{l}^\top (\log \mathbf{z} - \log \mathbf{u})\right) \left(\prod_{i=1}^d z_i^{-1}\right) 1_{\{\mathbf{z} \not\leq \mathbf{u}\mathbf{a}\}} \\
&\quad = \frac{C_{\mathbf{u}\mathbf{a}}(Q, \boldsymbol{l} + Q \log \mathbf{u})}{\exp\left(\frac{1}{2} \log \mathbf{u}^\top Q \log \mathbf{u} + \boldsymbol{l}^\top \log \mathbf{u}\right) C_{\mathbf{a}}(Q, \boldsymbol{l})} f_{\mathbf{u}\mathbf{a}}(\mathbf{z}; Q, \boldsymbol{l} + Q \log \mathbf{u})
\end{aligned}$$

This proves (i) as well as the equality

$$C_{\mathbf{u}\mathbf{a}}(Q, \boldsymbol{l} + Q \log \mathbf{u}) = \exp\left(\frac{1}{2} \log \mathbf{u}^\top Q \log \mathbf{u} + \boldsymbol{l}^\top \log \mathbf{u}\right) C_{\mathbf{a}}(Q, \boldsymbol{l}).$$

Using a similar reasoning, the change of variable $\tilde{\mathbf{z}} = \mathbf{z}^\beta$ yields

$$\mathbb{P}(\mathbf{Z}^\beta \in A) = \int_A f_{\mathbf{a}}(\tilde{\mathbf{z}}^{1/\beta}) \prod_{i=1}^d \frac{1}{\beta} \tilde{z}_i^{1/\beta - 1} \mathrm{d}\tilde{\mathbf{z}}$$

and simple computations show that

$$\begin{aligned}
&f_{\mathbf{a}}(\mathbf{z}^{1/\beta}) \prod_{i=1}^d \frac{1}{\beta} z_i^{1/\beta - 1} \\
&\quad = \frac{1}{C_{\mathbf{a}}(Q, \boldsymbol{l}) \beta^d} \exp\left(-\frac{1}{2} \log \mathbf{z}^\top \beta^{-1} Q \beta^{-1} \log \mathbf{z} + \boldsymbol{l}^\top \beta^{-1} \log \mathbf{z}\right) \left(\prod_{i=1}^d z_i^{-1}\right) 1_{\{\mathbf{z} \not\leq \mathbf{a}^\beta\}} \\
&\quad = \frac{C_{\mathbf{a}^\beta}(\beta^{-2}, \beta^{-1}\boldsymbol{l})}{C_{\mathbf{a}}(Q, \boldsymbol{l}) \beta^d} f_{\mathbf{a}^\beta}(\mathbf{z}; \beta^{-2} Q, \beta^{-1} \boldsymbol{l})
\end{aligned}$$

This implies (ii) as well as the equality

$$C_{a^\beta}(\beta^{-2} Q, \beta^{-1} \boldsymbol{l}) = \beta^d C_{\mathbf{a}}(Q, \boldsymbol{l}).$$

□

Remark 3.2. As a consequence of Proposition 3.1, Hüsler-Reiss Pareto vectors with $\mathbf{a} = \mathbf{1_d}$ and $\alpha = 1$ are particularly important, especially for simulation. Indeed, the random vector $\mathbf{Z} \rightsquigarrow \mathrm{HRPar}_{\mathbf{a}}(Q, \boldsymbol{l})$ with exponent $\alpha = -\boldsymbol{l}^\top \mathbf{1_d}$ satisfies $\mathbf{Z} \stackrel{d}{=} \mathbf{a}\widetilde{\mathbf{Z}}^{1/\alpha}$ where the random vector $\widetilde{\mathbf{Z}} \rightsquigarrow \mathrm{HRPar}_{\mathbf{1_d}}(\alpha^{-2} Q, \alpha^{-1}(\boldsymbol{l} - Q \log \mathbf{a}))$ takes values in $[0, \infty)^d \setminus [\mathbf{0}, \mathbf{1_d}]$ and has exponent 1.

Remark 3.3. The following equalities on the normalising constant seen in the proof of Proposition 3.1 are worth noting:

$$C_{\mathbf{ua}}(Q, \boldsymbol{l} + Q \log \mathbf{u}) = \exp\left(\frac{1}{2} \log \mathbf{u}^\top Q \log \mathbf{u} + \boldsymbol{l}^\top \log \mathbf{u}\right) C_{\mathbf{a}}(Q, \boldsymbol{l})$$

and

$$C_{\mathbf{a}^\beta}(\beta^{-2} Q, \beta^{-1} \boldsymbol{l}) = \beta^d C_{\mathbf{a}}(Q, \boldsymbol{l}).$$

As a consequence, we will often assume without loss of generality that $\mathbf{a} = \mathbf{1_d}$. The general case $\mathbf{a} \in (0, \infty)^d$ follows with the relation

$$C_{\mathbf{a}}(Q, \boldsymbol{l}) = \exp\left(-\frac{1}{2} \log \mathbf{a}^\top Q \log \mathbf{a} + \boldsymbol{l}^\top \log \mathbf{a}\right) C_{\mathbf{1_d}}(Q, \boldsymbol{l} - Q \log \mathbf{a}).$$

3.2.2 Exponential family properties

An important property of the Hüsler-Reiss Pareto distributions introduced above is to form an exponential family. Let E be an inner product space with dot product $\langle \cdot, \cdot \rangle$. A parametric family of densities $(f(\mathbf{z}; \theta))_{\theta \in \Theta}$ with $\Theta \subset E$ is a *canonical exponential family* if it can be written in the form

$$f(\mathbf{z}; \theta) = \frac{1}{C(\theta)} e^{\langle \theta, T(\mathbf{z}) \rangle} h(\mathbf{z}), \quad \mathbf{z} \in \mathbb{R}^d, \tag{3.1}$$

where $T: \quad \mathbb{R}^d \to E$ is the *natural sufficient statistic*. The exponential family is called a *full* exponential family if

$$\Theta = \left\{ t \in E \ : \ \int_{\mathbb{R}^d} e^{\langle t, T(\mathbf{z}) \rangle} h(\mathbf{z}) \, \mathrm{d}\mathbf{z} < \infty \right\}$$

is not contained in a strict subspace of E. For a detailed account on exponential family, the reader should refer to Barndorff-Nielsen [14].

Our main result in this section is the following Theorem.

Theorem 3.1. *Consider the $d(d+1)/2$-dimensional inner product space*

$$E = \{(A, \mathbf{b}) \in \mathbb{R}^{d \times d} \times \mathbb{R}^d \ : \ A^\top = A, \ A\mathbf{1_d} = \mathbf{0}\}$$

with inner product

$$\langle (A, \mathbf{a}), (A', \mathbf{a}') \rangle = \sum_{1 \le i,j \le d} A_{i,j} A'_{i,j} + \sum_{1 \le k \le d} a_k a'_k.$$

Define

$$\Theta = \left\{ (Q, \boldsymbol{l}) \in E \ : Q \text{ semi definite positive}, \ \operatorname{Ker} Q = \operatorname{span}(\mathbf{1_d}), \ \boldsymbol{l}^\top \mathbf{1_d} < 0 \right\}.$$

For all fixed $\mathbf{a} \in (0, \infty)^d$, the Hüsler-Reiss Pareto distributions $(f_{\mathbf{a}}(\mathbf{z}; \theta))_{\theta \in \Theta}$ form a full canonical exponential family with parameter $\theta = (Q, \boldsymbol{l}) \in \Theta$ and sufficient statistic

$$T(\mathbf{z}) = \left(-\frac{1}{2} \left(\log \mathbf{z} - \overline{\log \mathbf{z}} \right) \left(\log \mathbf{z} - \overline{\log \mathbf{z}} \right)^\top, \log \mathbf{z} \right), \tag{3.2}$$

with $\overline{\log \mathbf{z}} = d^{-1} (\mathbf{1_d}^\top \log z) \mathbf{1_d}$.

Proof. Without loss of generality, let $\mathbf{a} = \mathbf{1_d}$. Consider the intensity function

$$\tilde{\lambda}(\mathbf{z}) = \exp\left(-\frac{1}{2}\log\mathbf{z}^\top Q\log\mathbf{z} + \boldsymbol{l}^\top\log\mathbf{z}\right)\left(\prod_{i=1}^{d} z_i^{-1}\right), \quad \mathbf{z}\in(0,\infty)^d, \tag{3.3}$$

The symmetric matrix Q can be diagonalised in an orthonormal basis $Q = U\Delta U^\top$ with $\Delta = \mathrm{diag}(\lambda_1,\dots,\lambda_d)$ and U orthonormal. Thanks to the condition $Q\mathbf{1_d} = \mathbf{0}$, we can suppose $\lambda_1 = 0$ and the first column of U is equal to $\mathbf{U_1} = \mathbf{1_d}/\sqrt{d}$. Denote by Δ_{-1} (resp. v_{-1}) the matrix Δ (resp. vector v) with its first row and column removed (resp. first component removed), $\tilde{U}$ the $d\times(d-1)$ matrix obtained by removing the first column of U. The change of variable $\log\mathbf{z} = U\mathbf{v}$ gives

$$\int_{(0,\infty)^d} 1_{\{\mathbf{z}\not\leq\mathbf{a}\}}\tilde{\lambda}(\mathbf{z})\mathrm{d}\mathbf{z} = \int_{\mathbb{R}^d}\exp\left(-\frac{1}{2}\mathbf{v}_{-1}^\top\Delta_{-1}\mathbf{v}_{-1} + \boldsymbol{l}^\top\tilde{U}\mathbf{v}_{-1} + \boldsymbol{l}^\top U_1\mathbf{v}_1\right)1_{\mathbf{v}\in A}\mathrm{d}\mathbf{v}$$

where A equals

$$A = \left\{\mathbf{v}\in\mathbb{R}^d : v_1\mathbf{1_d}\not\leq -\tilde{U}\mathbf{v}_{-1}\right\} = \left\{\mathbf{v}\in\mathbb{R}^d : v_1 > a(\mathbf{v}_{-1}); a(v_{-1}) = \min_i -\sum_{j=1}^{d-1}\tilde{U}_{ij}v_{j+1}\right\}.$$

By Fubini theorem,

$$\int_{(0,\infty)^d} 1_{\{\mathbf{z}\not\leq\mathbf{a}\}}\tilde{\lambda}(\mathbf{z})\mathrm{d}\mathbf{z} = \int_{\mathbb{R}^{d-1}}\exp\left(-\frac{1}{2}\mathbf{v}_{-1}^\top\Delta_{-1}\mathbf{v}_{-1} + \boldsymbol{l}^\top\tilde{U}\mathbf{v}_{-1}\right)\int_{a(\mathbf{v}_{-1})}^{\infty}\exp\left(\boldsymbol{l}^\top\mathbf{U_1}v_1\right)\mathrm{d}v_1\mathrm{d}\mathbf{v}_{-1}.$$

The inner integral with respect to v_1 converges if and only if $\boldsymbol{l}^\top\mathbf{U}_1 < 0$ and then

$$\int_{(0,\infty)^d} 1_{\{\mathbf{z}\not\leq\mathbf{a}\}}\tilde{\lambda}(\mathbf{z})\mathrm{d}\mathbf{z} = \int_{\mathbb{R}^{d-1}}\exp\left(-\frac{1}{2}\mathbf{v}_{-1}^\top\Delta_{-1}\mathbf{v}_{-1} + \boldsymbol{l}^\top\tilde{U}\mathbf{v}_{-1} + \boldsymbol{l}^\top\mathbf{U_1}a(\mathbf{v}_{-1})\right)\mathrm{d}\mathbf{v}_{-1}$$

is finite if and only if Δ_{-1} is positive definite. This proves that the integral converge if and only if $(Q,\boldsymbol{l})\in\Theta$ and that the exponential family is full. □

In a general exponential model (3.1), the logarithm of the normalisation constant $C(\theta)$ is related to the cumulant generating function of the natural statistics T by the relation

$$\log\mathbb{E}_\theta\left[e^{\langle t,T(\mathbf{Z})\rangle}\right] = \log C(\theta+t) - \log C(\theta), \quad \theta,\theta+t\in\Theta.$$

If θ is an interior point of Θ, this implies

$$\mathbb{E}_\theta[T(\mathbf{Z})] = \frac{\partial\log C}{\partial\theta}(\theta) \quad\text{and}\quad \mathrm{Var}_\theta[T(\mathbf{Z})] = \frac{\partial^2\log C}{\partial\theta\partial\theta^\top}(\theta). \tag{3.6.1}$$

The computation of the normalisation constant $C(\theta)$ is hence particularly important.

Proposition 3.2. *In the Hüsler-Reiss Pareto model described in Theorem 3.1, we have*

$$C_{\mathbf{a}}(Q,\boldsymbol{l}) = (2\pi)^{(d-1)/2}\frac{1}{\alpha}\sum_{i=1}^{d} a_i^{-\alpha}\det(Q_{-i})^{-1/2}\exp\left\{\frac{1}{2}\boldsymbol{l}_{-i}^{\top}Q_{-i}^{-1}\boldsymbol{l}_{-i}\right\}\Phi_{d-1}\left(\log\frac{a_{-i}}{a_i};Q_{-i}^{-1}\boldsymbol{l}_{-i},Q_{-i}^{-1}\right),$$

where $\alpha = -\mathbf{1_d}^{\top}\boldsymbol{l}$*, the notation* $\boldsymbol{l}_{-i}$ *(resp.* $\mathbf{a}_{-i}$*) denotes the vector* $\boldsymbol{l}$ *(resp.* $\mathbf{a}$*) with its* i*th component removed,* Q_{-i} *the matrix* Q *with its* i*th column and row removed and* $\Phi_d(\mathbf{z};\mathbf{m},\Sigma)$ *denotes the cumulative distributive function at* $\mathbf{z}$ *of a* d*-dimensional multivariate Gaussian distribution with mean* $\mathbf{m}$ *and covariance* Σ*.*

The expression for $C_{\mathbf{a}}(Q,\boldsymbol{l})$ was first established by Huser and Davison [108]. We provide here a direct proof that will be needed for further reference (proof of Proposition 3.3).

Proof. With $\tilde{\lambda}$ the function defined by Equation (3.3), the normalisation constant $C_{\mathbf{a}}(Q,\boldsymbol{l})$ is given by

$$C_{\mathbf{a}}(Q,\boldsymbol{l}) = \int_{[\mathbf{0},\mathbf{a}]^c}\tilde{\lambda}(\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

Since $[\mathbf{0},\mathbf{a}]^c = \cup_{i=1}^{d}A_i$ with

$$A_i = \left\{\mathbf{z}\in\mathbb{R}^d\ :\ z_i > a_i, z_{-i}/z_i \le a_{-i}/a_i\right\},\quad i=1,\dots,d,$$

we have

$$C_{\mathbf{a}}(Q,\boldsymbol{l}) = \sum_{i=1}^{d}\int_{A_i}\tilde{\lambda}(\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

Using the homogeneity relation (2.7) with $x = z_i$, we get

$$\tilde{\lambda}(\mathbf{z}) = \tilde{\lambda}(z_i\,\mathbf{z}/z_i) = z_i^{-d-\alpha}\tilde{\lambda}(\mathbf{z}/z_i).$$

Since the ith component of $\mathbf{z}/z_i$ is equal to 1, we have also

$$\tilde{\lambda}(\mathbf{z}/z_i) = \exp\left(-\frac{1}{2}\log\tilde{\mathbf{z}}_{-i}^{\top}Q_{-i}\log\tilde{\mathbf{z}}_{-i} + \boldsymbol{l}_{-i}^{\top}\log\tilde{\mathbf{z}}_{-i}\right)\prod_{j\neq i}^{d}\tilde{z}_j^{-1},\quad \tilde{\mathbf{z}}_{-i} = \mathbf{z}_{-i}/z_i.$$

These relations imply

$$\begin{aligned}
&\int_{A_i}\tilde{\lambda}(\mathbf{z})\mathrm{d}\mathbf{z}\\
&=\int_{(0,\infty)^d}1_{\{z_i>a_i,\ \mathbf{z}_{-i}/z_i\le\mathbf{a}_{-i}/a_i\}}z_i^{-d-\alpha}\tilde{\lambda}(\mathbf{z}/z_i)\,\mathrm{d}\mathbf{z}\\
&=\int_{(0,\infty)^d}1_{\{z_i>a_i,\ \tilde{\mathbf{z}}_{-i}\le\mathbf{a}_{-i}/a_i\}}z_i^{-d-\alpha}\exp\left(-\frac{1}{2}\log\tilde{\mathbf{z}}_{-i}^{\top}Q_{-i}\log\tilde{\mathbf{z}}_{-i}+\boldsymbol{l}_{-i}^{\top}\log\tilde{\mathbf{z}}_{-i}\right)\prod_{j\neq i}^{d}\tilde{z}_j^{-1}\,\mathrm{d}\mathbf{z}\\
&=\int_{a_i}^{\infty}\int_{[0,\mathbf{a}_{-i}/a_i]}z_i^{-\alpha-1}\exp\left(-\frac{1}{2}\log\tilde{\mathbf{z}}_{-i}^{\top}Q_{-i}\log\tilde{\mathbf{z}}_{-i}+\boldsymbol{l}_{-i}^{\top}\log\tilde{\mathbf{z}}_{-i}\right)\left(\prod_{j\neq i}^{d}\tilde{z}_j^{-1}\right)\mathrm{d}z_i\mathrm{d}\tilde{\mathbf{z}}_{-i}\qquad(3.4)\\
&=\frac{1}{\alpha}a_i^{-\alpha}\int_{[0,\mathbf{a}_{-i}/a_i]}\exp\left(-\frac{1}{2}\log\tilde{\mathbf{z}}_{-i}^{\top}Q_{-i}\log\tilde{\mathbf{z}}_{-i}+\boldsymbol{l}_{-i}^{\top}\log\tilde{\mathbf{z}}_{-i}\right)\left(\prod_{j\neq i}^{d}\tilde{z}_j^{-1}\right)\mathrm{d}\tilde{\mathbf{z}}_{-i}
\end{aligned}$$

where we have used for the third inequality the change of variable $\mathbf{z} \to (z_i, \tilde{\mathbf{z}}_{-i})$. In the last integral with respect to $\tilde{\mathbf{z}}_{-i}$, we recognise a log-normal density (up to a multiplicative factor), so that

$$\int_{[0,\mathbf{a}_{-i}/a_i]} \exp\left(-\frac{1}{2}\log\tilde{\mathbf{z}}_{-i}^\top Q_{-i}\log\tilde{\mathbf{z}}_{-i} + \boldsymbol{l}_{-i}^\top \log\tilde{\mathbf{z}}_{-i}\right)\left(\prod_{j\neq i}^d \tilde{z}_j^{-1}\right)\mathrm{d}\tilde{\mathbf{z}}_{-i}$$
$$=(2\pi)^{(d-1)/2}\det(Q_{-i})^{-1/2}\exp\left\{\frac{1}{2}\boldsymbol{l}_{-i}^\top Q_{-i}^{-1}\boldsymbol{l}_{-i}\right\}\Phi_{d-1}\left(\log(\mathbf{a}_{-i}/a_i); Q_{-i}^{-1}\boldsymbol{l}_{-i}, Q_{-i}^{-1}\right).$$

The result follows:

$$C_{\mathbf{a}}(Q,\boldsymbol{l}) = \sum_{i=1}^d \int_{A_i} \tilde{\lambda}(\mathbf{z})\,\mathrm{d}\mathbf{z}$$
$$= (2\pi)^{(d-1)/2}\frac{1}{\alpha}\sum_{i=1}^d a_i^{-\alpha}\det(Q_{-i})^{-1/2}\exp\left\{\frac{1}{2}\boldsymbol{l}_{-i}^\top Q_{-i}^{-1}\boldsymbol{l}_{-i}\right\}\Phi_{d-1}(\log(\mathbf{a}_{-i}/a_i); Q_{-i}^{-1}\boldsymbol{l}_{-i}, Q_{-i}^{-1}).$$

□

Corollary 3.1. *Let* $\mathbf{Z} \rightsquigarrow \mathrm{HRPar}_{\mathbf{a}}(Q,\boldsymbol{l})$ *with exponent* $\alpha = -\boldsymbol{l}^\top\mathbf{1_d} > 0$. *Then,*

(i) for all $\mathbf{u} = (u_1,\dots,u_d)$ *such that* $\sum_{i=1}^d u_i < \alpha$, *we have*

$$\mathbb{E}\left[\prod_{i=1}^d Z_i^{u_i}\right] = \frac{C_{\mathbf{a}}(Q,\boldsymbol{l}+\mathbf{u})}{C_{\mathbf{a}}(Q,\boldsymbol{l})}.$$

(ii) The expectation and covariance matrix of $\log\mathbf{Z}$ *are given by*

$$\mathbb{E}\left[\log Z_i\right] = \frac{\partial\log C_{\mathbf{a}}}{\partial l_i}(Q,\boldsymbol{l}) \quad i=1,\dots,d,$$

and

$$\mathrm{Cov}\left(\log Z_i, \log Z_j\right) = \frac{\partial^2\log C_{\mathbf{a}}}{\partial l_i\partial l_j}(Q,\boldsymbol{l}) \quad i,j=1,\dots,d.$$

(iii) Moreover, the expectation and covariance matrix of $\log\mathbf{Z}$ *satisfies*

$$\mathbb{E}[(\log\mathbf{Z} - \overline{\log\mathbf{Z}})(\log\mathbf{Z} - \overline{\log\mathbf{Z}})^\top] = \frac{\partial\log C_{\mathbf{a}}}{\partial Q}(Q,\boldsymbol{l})$$

Proof. For $\theta = (Q,\boldsymbol{l}) \in \Theta$, we have for all $\mathbf{u} = (u_1,\dots,u_d)$ such that $\sum_{i=1}^d u_i < \alpha$

$$\theta + (0,\mathbf{u}) \in \Theta$$

by definition of α. Using equality (3.6.1) with $t = (0,\mathbf{u})$, we have

$$\log\mathbb{E}_\theta\left[e^{\langle\mathbf{u},\log\mathbf{Z}\rangle}\right] = \log C(Q,\boldsymbol{l}+\mathbf{u}) - \log C(Q,\boldsymbol{l}),$$

taking to the exponential and developing the product implies (i).
The results (ii) and (iii) are straightforwards applications of (3.6.1). □

Example 3.2. *In dimension $d = 2$ with $a = (1,1)$ and the same notations as in Example* (3.1)*, we have*

$$C(Q,\boldsymbol{l}) = \frac{\sqrt{2\pi}}{\alpha\sqrt{c}}\left\{e^{l_1^2/2c}\Phi\left(-l_1/\sqrt{c}\right) + e^{l_2^2/2c}\Phi\left(-l_2/\sqrt{c}\right)\right\}.$$

The first order partial derivatives of $\log C$ *are equal to*

$$\begin{aligned}
\frac{\partial \log C}{\partial l_1} &= -\frac{1}{l_1+l_2} + \frac{cl_1\Phi\left(-l_1/\sqrt{c}\right) - \varphi(-l_1/\sqrt{c})/\sqrt{c}}{\Phi\left(-l_1/\sqrt{c}\right) + e^{c(l_2^2-l_1^2)/2}\Phi\left(-l_2/\sqrt{c}\right)}\\
\frac{\partial \log C}{\partial l_2} &= -\frac{1}{l_1+l_2} + \frac{cl_2\Phi\left(-l_2/\sqrt{c}\right) - \varphi(-l_2/\sqrt{c})/\sqrt{c}}{\Phi\left(-l_2/\sqrt{c}\right) + e^{c(l_1^2-l_2^2)/2}\Phi\left(-l_1/\sqrt{c}\right)}\\
\frac{\partial \log C}{\partial c} &= -\frac{1}{2c} + \frac{1}{2}\frac{l_1\Phi(-l_1/\sqrt{c}) - l_1c^{-3/2}\phi(-l_1/\sqrt{c})}{\Phi\left(-l_1/\sqrt{c}\right) + e^{c(l_2^2-l_1^2)/2}\Phi\left(-l_2/\sqrt{c}\right)}\\
&\quad + \frac{1}{2}\frac{l_2\Phi(-l_2/\sqrt{c}) - l_2c^{-3/2}\phi(-l_2/\sqrt{c})}{\Phi\left(-l_2/\sqrt{c}\right) + e^{c(l_1^2-l_2^2)/2}\Phi\left(-l_1/\sqrt{c}\right)}
\end{aligned}$$

This formulas provides respectively the expectations $\mathbb{E}[\log Z_1]$*,* $\mathbb{E}[\log Z_2]$ *and* $-\frac{1}{8}\mathbb{E}[(\log Z_1 - \log Z_2)^2]$*. Formulas for the general case* $\mathbf{a} = (a_1, a_2)$ *can be deduced using Proposition 3.1.*

3.2.3 Simulation of HR-Pareto random vectors

We now consider the simulation of an Hüsler-Reiss Pareto random vector $\mathbf{Z} \rightsquigarrow \mathrm{HRPar}_{\mathbf{a}}(Q,\boldsymbol{l})$. Thanks to the transformation property (3.1), we focus on the case $\mathbf{a} = \mathbf{1_d}$. In the following proposition, we denote by $S = \{\mathbf{x} \in (0,\infty)^d : \|\mathbf{x}\|_\infty = 1\}$ the unit sphere and we use $S = \cup_{i=1}^d S_i$ with $S_i = \{\mathbf{x} \in S : x_i = 1\}$.

Proposition 3.3. *Let* $\mathbf{Z} \rightsquigarrow \mathrm{HRPar}_{1_d}(Q,\boldsymbol{l})$ *with exponent* $\alpha > 0$*. Then* $R = \|\mathbf{Z}\|$ *and* $\boldsymbol{\Theta} = \mathbf{Z}/\|\mathbf{Z}\|$ *are independent and such that*

- *R is a* Pareto(α)*-distributed real random variable, i.e.,* $\mathbb{P}(R > r) = r^{-\alpha}$*,* $r > 1$*;*
- *$\boldsymbol{\Theta}$ is a random vector on S satisfying, for $i = 1,\dots,d$,*

$$\mathbb{P}(\boldsymbol{\Theta} \in S_i) = \frac{\det(Q_{-i})^{-1/2}\exp\left\{\frac{1}{2}\boldsymbol{l}_{-i}^\top Q_{-i}^{-1}\boldsymbol{l}_{-i}\right\}\Phi_{d-1}\left(\mathbf{0}; Q_{-i}^{-1}\boldsymbol{l}_{-i}, Q_{-i}^{-1}\right)}{\sum_{j=1}^d \det(Q_{-j})^{-1/2}\exp\left\{\frac{1}{2}\boldsymbol{l}_{-j}^\top Q_{-j}^{-1}\boldsymbol{l}_{-j}\right\}\Phi_{d-1}\left(\mathbf{0}; Q_{-j}^{-1}\boldsymbol{l}_{-j}, Q_{-j}^{-1}\right)} \tag{3.5}$$

and, given $\boldsymbol{\Theta} \in S_i$, $\Theta_i = 1$ and

$$\mathcal{L}(\boldsymbol{\Theta}_{-i} \mid \boldsymbol{\Theta} \in S_i) = \mathcal{L}(\exp(\mathbf{G}_i) \mid \mathbf{G}_i \le 0) \quad \textit{with} \quad \mathbf{G}_i \rightsquigarrow \mathcal{N}_{d-1}(Q_{-i}^{-1}\boldsymbol{l}_{-i}, Q_{-i}^{-1}).$$

Proof. The proof is mostly a reinterpretation of the computations from the proof of Proposition 3.2. The density of $\mathbf{Z} \rightsquigarrow \mathrm{HRPar}_{\mathbf{1_d}}(Q,\boldsymbol{l})$ is given by

$$f_{\mathbf{1_d}}(\mathbf{z}; Q, \boldsymbol{l}) = \frac{1}{C_{\mathbf{1_d}}(Q,\boldsymbol{l})}1_{\{\|\mathbf{z}\|>1\}}\tilde{\lambda}(\mathbf{z}), \quad \mathbf{z} \in (0,\infty)^d,$$

with $\tilde{\lambda}$ the function defined by Equation (3.3). From the proof of Proposition 3.2, we have

$$C_{\mathbf{1_d}}(Q, \boldsymbol{l}) = \sum_{i=1}^{d} \int_{A_i} \tilde{\lambda}(\mathbf{z}) \mathrm{d}\mathbf{z}$$

with

$$A_i = \{\mathbf{z} \in (0, \infty)^d \,:\, \|\mathbf{z}\| > 1, \mathbf{z}/\|\mathbf{z}\| \in S_i\}, \quad i = 1, \ldots, d.$$

The expression for A_i here is slightly different but equivalent since $\mathbf{a} = \mathbf{1_d}$. Consequently, we get

$$\mathbb{P}(\boldsymbol{\Theta} \in S_i) = \int_{(0,\infty)^d} 1_{\{\mathbf{z}/\|\mathbf{z}\| \in S_i\}} f_{\mathbf{1_d}}(\mathbf{z}; Q, \boldsymbol{l}) \mathrm{d}\mathbf{z} = \frac{1}{C_{\mathbf{1_d}}(Q, \boldsymbol{l})} \int_{A_i} \tilde{\lambda}(\mathbf{z}) \mathrm{d}\mathbf{z}$$

which yields Equation (3.5) in view of Proposition 3.2 and its proof.

On the other hand, when $\mathbf{Z} \in A_i$ or equivalently $\boldsymbol{\Theta} \in S_i$, we have $R = \|\mathbf{Z}\| = Z_i$ whence the change of variable $\mathbf{z} \to (z_i, \tilde{\mathbf{z}}_{-i})$ in Equation (3.4) provides exactly the joint distribution of $(R, \boldsymbol{\Theta}_{-i})$. This amounts to be the product of α-Pareto and log-normal distributions, proving the independence of R and $\boldsymbol{\Theta}$ and the form of their distribution. □

In order to simulate the Gaussian random vector $\mathbf{G}_i$ conditioned on $\mathbf{G}_i \leq 0$, we propose a recursive sampling procedure. Let $i \in \{1, \ldots, d\}$ be fixed and denote by $G_{i,j}$ the components of G_i. We first set $G_{i,i} = 0$ and $J = \{1, \ldots, d\} \setminus \{i\}$ the set of indices to sample. For $j \in J$, the conditional distribution of $G_{i,j}$ given the already sampled components $\mathbf{G_{i,J^c}}$ has a Gaussian distribution with mean and variance

$$m_{i,j} = \Big(Q_{J,J}^{-1}\big(\boldsymbol{l}_J - Q_{J,J^c}\mathbf{G_{i,J^c}}\big)\Big)_j \quad \text{and} \quad \sigma_{i,j}^2 = \Big(Q_{J,J}^{-1}\Big)_{j,j} \tag{3.6}$$

subject to the constraint $G_{i,j} \leq 0$. By the inversion method, we can sample from $G_{i,j}$ as

$$G_{i,j} = m_{i,j} + \sigma_{i,j}\Phi^{-1}\Big(\Phi(-m_{i,j}/\sigma_{i,j})U_j\Big), \quad U_j \rightsquigarrow \mathrm{Unif}([0,1]),$$

where Φ denotes the standard normal cumulative distribution function. Then, we replace J by $J \setminus \{j\}$ and repeat the procedure with the next component to sample until J is empty.

Note that the above computations are closely related to the distribution of extremal functions in the conditional sampling procedure of the Brown-Resnick max-stable process, see Dombry et al. [69, section 2.2]. Based on Proposition 3.3 and the above recursive scheme, Algorithm 2 describes a simulation procedure for Hüsler-Reiss Pareto random vectors.

3.2.4 Maximum likelihood inference

The exponential family property of the Hüsler-Reiss Pareto distributions makes maximum likelihood inference particularly convenient. We always suppose the threshold $\mathbf{a} \in (0, \infty)^d$ to be known and estimate the parameter $\theta = (Q, \boldsymbol{l}) \in \Theta$ from observations $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(n)} \in$

Input: the parameters Q and l of the HR-Pareto distribution
Output: a sample $Z \rightsquigarrow \mathrm{HRPar}_{1_d}(Q, \boldsymbol{l})$
Compute $\alpha = -l^\top 1_d$ and sample $R \rightsquigarrow \mathrm{Pareto}(\alpha)$.
Compute $p_i = P(\Theta \in S_i)$, $i = 1, \dots, d$, according to Eq. (3.5).
Sample i from the distribution $(p_1, \dots, p_d)$ and set $J = \{1, \dots, d\} \setminus \{i\}$.
Initialise $G = 0_d$ (d-dimensional null vector).
for $j \in J$ **do**
  Compute m, σ^2 according to Eq. (3.6).
  Sample $U \rightsquigarrow \mathrm{Unif}([0,1])$ and set $G_j = m + \sigma\Phi^{-1}\Big(\Phi(-m/\sigma)U\Big)$.
  Set $J = J \setminus \{j\}$.
end
Set $\Theta = \exp(G)$ and $Z = R\Theta$.
return Z.

Algorithm 2: Simulation of a Hüsler-Reiss Pareto random vector

$(0,\infty)^d \setminus [\mathbf{0}, \mathbf{a}]$. In the Hüsler-Reiss Pareto model, the log-likelihood of the sample writes, for $\theta = (Q, \boldsymbol{l}) \in \Theta$,

$$\begin{aligned} L_n(\theta; \mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(n)}) &= \frac{1}{n}\sum_{i=1}^n \log f_{\mathbf{a}}(\mathbf{z}^{(i)}; Q, \boldsymbol{l}) \\ &= \langle (Q, \boldsymbol{l}), \overline{T}_n \rangle - \log C_a(Q, \boldsymbol{l}) + C \end{aligned}$$

where $\overline{T}_n$ is the sufficient statistic defined by

$$\overline{T}_n = \left(-\frac{1}{2n}\sum_{i=1}^n (\log \mathbf{z}^{(i)} - \overline{\log \mathbf{z}^{(i)}})(\log \mathbf{z}^{(i)} - \overline{\log \mathbf{z}^{(i)}})^\top, \frac{1}{n}\sum_{i=1}^n \log \mathbf{z}^{(i)} \right)$$

and the constant term C does not depend on the parameter $\theta = (Q, \boldsymbol{l})$. Using the classical theory of maximum likelihood estimation for exponential families, we obtain the following result, regarding existence, uniqueness and asymptotic normality of the maximum likelihood estimator

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta \in \Theta} L_n(\theta; \mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(n)}).$$

Theorem 3.2. *Let* $\mathbf{a} \in (0,\infty)^d$ *and* $n \geq 1$.

(i) (existence and uniqueness) For observations $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(n)} \in [\mathbf{0}, \mathbf{a}]^c$*, the log-likelihood* $(Q, \boldsymbol{l}) \mapsto L_n(Q, \boldsymbol{l}; \mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(n)})$ *is strictly concave on* Θ*. A maximum likelihood estimator exists if and only the sample covariance matrix*

$$V_n = \frac{1}{n}\sum_{i=1}^n \log \mathbf{z}^{(i)} \log \mathbf{z}^{(i)\top} - \left(\frac{1}{n}\sum_{i=1}^n \log \mathbf{z}^{(i)}\right)\left(\frac{1}{n}\sum_{i=1}^n \log \mathbf{z}^{(i)}\right)^\top$$

is conditionally definite positive in the sense that $\mathbf{v}^\top V_n \mathbf{v} > 0$ *for all* $\mathbf{v} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ *such that* $\mathbf{v}^\top \mathbf{1_d} = 0$. *If it exists, the maximum likelihood* $\hat{\theta}_n^{mle}$ *estimator is the unique solution to the score equation*

$$\frac{\partial \log C_a}{\partial \theta}(\theta) = \overline{T}_n, \quad \theta \in \Theta. \tag{3.7}$$

(ii) (asymptotic normality) Let $\theta = (Q, \boldsymbol{l}) \in \Theta$ *and assume* $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(n)}$ *are generated from the distribution* $\mathrm{HRPar}_{\mathbf{a}}(Q, \boldsymbol{l})$. *Then, for* $n \geq d-1$, *there exists almost surely a unique maximum likelihood estimator* $\hat{\theta}_n^{mle}$ *which is asymptotically normal and efficient, that is*

$$\sqrt{n}(\hat{\theta}_n^{mle} - \theta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, I(\theta)^{-1}), \quad \text{as } n \to \infty,$$

where $I(\theta)$ *is the Fisher Information matrix given by*

$$I(\theta) = -\frac{\partial^2 \log C_{\mathbf{a}}}{\partial \theta \partial \theta^\top}(\theta).$$

Remark 3.4. *In statement i), if* $\mathbf{1_d}^\top \log \mathbf{z}^{(1)}, \ldots, \mathbf{1_d}^\top \log \mathbf{z}^{(n)}$ *are not all equal for* $i = 1, \cdots, n$ *then the condition* V_n *conditionally definite positive is equivalent to* V_n *definite positive.*

The proof of Theorem 3.2 relies on the following Lemma.

Lemma 3.1. *Recall the definition* (3.1) *of the sufficient statistic* $T(\mathbf{z})$. *Then, the closed convex hull of the set*

$$S = \left\{ T(\mathbf{z}) \; ; \; \mathbf{z} \in (0, \infty)^d, \; \mathbf{z} \not\leq 1_d \right\}$$

is equal to

$$C = \left\{ (Q, \boldsymbol{l}) \in E : Q \preceq -\frac{1}{2}(\boldsymbol{l} - \bar{\boldsymbol{l}})(\boldsymbol{l} - \bar{\boldsymbol{l}})^\top \right\},$$

where $Q_1 \preceq Q_2$ *means that the symmetric matrix* $Q_2 - Q_1$ *is semi-definite positive.*

Proof of Lemma 3.1. The change of variable $\mathbf{u} = \log \mathbf{z}$ shows that

$$S = \left\{ \left(-\frac{1}{2}(\mathbf{u} - \overline{\mathbf{u}})(\mathbf{u} - \overline{\mathbf{u}})^\top, \mathbf{u} \right) \; , \mathbf{u} \not\leq 0 \right\}$$

where $\overline{\mathbf{u}} = d^{-1}(\mathbf{1_d}^\top \log \mathbf{z})\mathbf{1_d}$. It is easily shown that C is closed, convex and contains S, so that $\overline{\mathrm{conv}}(S) \subset \overline{\mathrm{conv}}(C) = C$. We consider now the reverse inclusion. Consider $\mathbf{U}, \mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots$ i.i.d. with mean $\boldsymbol{l}$, variance Σ and such that $\mathbf{U} \not\leq 0$ a.s. The random element

$$S_n = \frac{1}{n} \sum_{i=1}^{n} \left(-\frac{1}{2}(\mathbf{U}^{(i)} - \overline{\mathbf{U}^{(i)}})(\mathbf{U}^{(i)} - \overline{\mathbf{U}^{(i)}})^\top, \mathbf{U}^{(i)} \right).$$

belongs to $\mathrm{conv}(S)$ and, by the law of large numbers,

$$S_n \xrightarrow{a.s.} S_\infty = \left(-\frac{1}{2}\mathbb{E}\left((\mathbf{U} - \overline{\mathbf{U}})(\mathbf{U} - \overline{\mathbf{U}})^\top \right), \mathbb{E}(\mathbf{U}) \right), \quad n \to \infty,$$

so that $S_\infty \in \overline{\text{conv}}(S)$. We prove below that for all $(Q, \boldsymbol{l}) \in C$, one can choose Σ such that $S_\infty = (Q, \boldsymbol{l}) \in \overline{\text{conv}}(S)$, proving the reverse inclusion $C \subset \overline{\text{conv}}(S)$. Using $\overline{\mathbf{U}} = d^{-1}\mathbf{1_d}\mathbf{1_d}^\top\mathbf{U}$, we deduce

$$\begin{aligned}
\mathbb{E}\left((\mathbf{U}-\overline{\mathbf{U}})(\mathbf{U}-\overline{\mathbf{U}})^\top\right) &= \mathbb{E}\left((\mathbf{U}-d^{-1}\mathbf{1_d}\mathbf{1_d}^\top\mathbf{U})(\mathbf{U}-d^{-1}\mathbf{1_d}\mathbf{1_d}^\top\mathbf{U})^\top\right) \\
&= \mathbb{E}\left(\left(I-\frac{1}{d}\mathbf{1_d}\mathbf{1_d}^\top\right)\mathbf{U}\mathbf{U}^\top\left(I-\frac{1}{d}\mathbf{1_d}\mathbf{1_d}^\top\right)^\top\right) \\
&= \left(I-\frac{1}{d}\mathbf{1_d}\mathbf{1_d}^\top\right)(\Sigma+\boldsymbol{l}\boldsymbol{l}^\top)\left(I-\frac{1}{d}\mathbf{1_d}\mathbf{1_d}^\top\right)^\top
\end{aligned}$$

It is proved in Lemma 3.2 that the linear operator on the space of symmetric $d \times d$ matrices defined by

$$P : M \mapsto \left(I-\frac{1}{d}\mathbf{1_d}\mathbf{1_d}^\top\right) M \left(I-\frac{1}{d}\mathbf{1_d}\mathbf{1_d}^\top\right)^\top$$

is the orthogonal projection on the linear subspace $\{M : M\mathbf{1_d} = \mathbf{0}\}$. Therefore, for Σ such that $P(\Sigma + \boldsymbol{l}\boldsymbol{l}^\top) = Q$, we have

$$S_n \longrightarrow (P(\Sigma + \boldsymbol{l}\boldsymbol{l}^\top), \boldsymbol{l}) = (Q, \boldsymbol{l}).$$

In particular, since we can take Σ among all symmetric positive semi-definite matrix, the choice $\Sigma = -2Q - P(\boldsymbol{l}\boldsymbol{l}^\top)$ which is positive by definition of C leads to the result. Therefore $C \subset \overline{\text{conv}(S)}$. □

Proof of Theorem 3.2. We assume here without loss of generality that $\mathbf{a} = \mathbf{1_d}$. The cumulant transform $\theta \in (Q, \boldsymbol{l}) \in \Theta \mapsto \log C_{\mathbf{a}}(Q, \boldsymbol{l})$ is a strictly convex function. Therefore the log-likelihood L_n is strictly concave as a difference of a linear function and a strictly convex function. The general theory for exponential families (see e.g. Barndorff-Nielsen [14, Theorem 9.13]) ensures that the maximum likelihood estimator exists if and only if the sufficient statistic $\overline{T}_n$ belongs to the interior of the closed convex hull of the support of T, that is $\overline{T}_n \in \text{int}(\overline{\text{conv}(S)}) = \text{int}(C)$ with S and C defined in Lemma 3.1. In this case, Theorem 9.13 in Barndorff-Nielsen [14] implies that the maximum likelihood estimator is unique and solves the score equation (3.7). So in order to prove statement (i), it remains to prove that $\overline{T}_n \in \text{int}(\overline{\text{conv}(S)})$ if and only if V_n is conditionally definite positive. Note that, by Lemma 3.1,

$$\text{int}(\overline{\text{conv}(S)}) = \text{int}(C) = \left\{(Q, \boldsymbol{l}) \in E \ : \ Q \prec -\frac{1}{2}(\boldsymbol{l}-\bar{\boldsymbol{l}})(\boldsymbol{l}-\bar{\boldsymbol{l}})^\perp \text{ on } \text{span}(\mathbf{1_d})^\top\right\},$$

where $Q_1 \prec Q_2$ on $\text{span}(\mathbf{1_d})^\top$ means that $\mathbf{v}^\top(Q_2 - Q_1)\mathbf{v} > 0$ for all $\mathbb{R}^d \setminus \{\mathbf{0}\}$ such that

$\mathbf{v}^\top \mathbf{1_d} = 0$. For such $\mathbf{v}$ and for $(Q, \boldsymbol{l}) = \overline{T}_n$, we have

$$\begin{aligned}
&\mathbf{v}^\top \left(-Q - \frac{1}{2}(\boldsymbol{l} - \bar{\boldsymbol{l}})(\boldsymbol{l} - \bar{\boldsymbol{l}})^\top \right) \mathbf{v} \\
=&\mathbf{v}^\top \left(\frac{1}{2n} \sum_{i=1}^n (\log \mathbf{z}^{(i)})(\log \mathbf{z}^{(i)})^\top - \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^n \log \mathbf{z}^{(i)} \right) \left(\frac{1}{n} \sum_{i=1}^n \log \mathbf{z}^{(i)} \right)^\top \right) \mathbf{v} \geq 0 \\
=&\mathbf{v}^\top V_n \mathbf{v}
\end{aligned}$$

whence we deduce that $\overline{T}_n \in \mathrm{int}(\overline{\mathrm{conv}(S)})$ if and only if V_n is conditionally positive.

Statement *ii*) follows directly from the general theory of exponential families since the Hüsler-Reiss distributions form a full rank exponential family (see e.g. Van der Vaart [176, Theorem 4.6]).

□

3.3 The generalised Hüsler-Reiss Pareto model

3.3.1 Definition and transformation properties

Definition 3.2. *Let $d \geq 2$ and define Θ the set of all $\theta = (\boldsymbol{\alpha}, Q, \boldsymbol{l})$ such that:*

- $\boldsymbol{\alpha} \in (0, \infty)^d$,
- $Q \in \mathbb{R}^{d \times d}$ *is symmetric semi-definite positive and* $\mathrm{Ker} Q = \mathrm{span}(\mathbf{1_d})$,
- $\boldsymbol{l} \in \mathbb{R}^d$ *satisfies* $\boldsymbol{l}^\top \mathbf{1_d} = -1$.

For $\mathbf{a} \in (0, \infty)^d$, the generalised Hüsler-Reiss Pareto model on $[\mathbf{0}, \mathbf{a}]^c = [0, \infty)^d \setminus [\mathbf{0}, \mathbf{a}]$ with parameters $\theta = (\boldsymbol{\alpha}, Q, \boldsymbol{l})$ is defined by the density

$$f_{\mathbf{a}}(\mathbf{z}; \theta) = \frac{1}{C_{\mathbf{a}}(\theta)} \exp\left(-\frac{1}{2} \log \mathbf{z}^\top D_{\boldsymbol{\alpha}} Q D_{\boldsymbol{\alpha}} \log \mathbf{z} + \boldsymbol{l}^\top D_{\boldsymbol{\alpha}} \log \mathbf{z} \right) \left(\prod_{i=1}^d z_i^{-1} \right) 1_{\{\mathbf{z} \not\leq \mathbf{a}\}} \tag{3.8}$$

with $C_{\mathbf{a}}(\theta)$ the normalisation constant and $D_{\boldsymbol{\alpha}}$ the diagonal matrix with diagonal $\boldsymbol{\alpha}$.
We write $\mathbf{Z} \rightsquigarrow \mathrm{HRPar}_{\mathbf{a}}(\boldsymbol{\alpha}, Q, \boldsymbol{l})$ for a random vector $\mathbf{Z}$ with density $f_{\mathbf{a}}(\mathbf{z}; \boldsymbol{\alpha}, Q, \boldsymbol{l})$.

For $\lambda > 0$, the substitution $(\boldsymbol{\alpha}, Q, \boldsymbol{l}) \mapsto (\lambda \boldsymbol{\alpha}, \lambda^{-1/2} Q, \lambda^{-1} \boldsymbol{l})$ leaves Equation (3.8) invariant so that the condition $\boldsymbol{l}^\top \mathbf{1_d} = -1$ is meant to ensure that the model is identifiable. In the case $\boldsymbol{\alpha} = \bar{\alpha} \mathbf{1_d}$ with $\bar{\alpha} > 0$, the generalised Hüsler-Reiss model coincides with the Hüsler-Reiss Pareto model since $f_{\mathbf{a}}(\mathbf{z}; \alpha, Q, \boldsymbol{l}) = f_{\mathbf{a}}(\mathbf{z}; \bar{\alpha}^2 Q, \bar{\alpha} \boldsymbol{l})$ and $\bar{\alpha}$ is the tail index.

Similarly as HR-Pareto distributions, generalised HR-Pareto distributions enjoy a stability property under scale and power transformations.

Proposition 3.4. *Let $\mathbf{Z} \rightsquigarrow \mathrm{HRPar}_{\mathbf{a}}(\boldsymbol{\alpha}, Q, \boldsymbol{l})$.*

(i) For all $\mathbf{u} \in (0,\infty)^d$, $\mathbf{uZ} \rightsquigarrow \mathrm{HRPar}_{\mathbf{ua}}(\boldsymbol{\alpha}, Q, \boldsymbol{l} + QD_{\boldsymbol{\alpha}} \log \mathbf{u})$.

(ii) For all $\boldsymbol{\beta} \in (0,\infty)^d$, $\mathbf{Z}^{\boldsymbol{\beta}} \rightsquigarrow \mathrm{HRPar}_{\mathbf{a}^{\boldsymbol{\beta}}}(\boldsymbol{\alpha}/\boldsymbol{\beta}, Q, \boldsymbol{l})$.

Proof. The change of variable $\tilde{\mathbf{z}} = \mathbf{uz}$ implies

$$\mathbb{P}(\mathbf{uZ} \in A) = \int_A f_{\mathbf{a}}(\tilde{\mathbf{z}}/\mathbf{u}; \boldsymbol{\alpha}, Q, \boldsymbol{l}) \prod_{i=1}^d u_i^{-1} \mathrm{d}\tilde{\mathbf{z}}.$$

Similarly as in the proof of Proposition (3.1), we check that

$$\begin{aligned} & f_{\mathbf{a}}(\mathbf{z}/\mathbf{u}; \boldsymbol{\alpha}, Q, \boldsymbol{l}) \prod_{i=1}^d u_i^{-1} \\ & \quad = \frac{C_{\mathbf{ua}}(\boldsymbol{\alpha}, Q, \boldsymbol{l} + Q \log \mathbf{u})}{\exp\left\{\frac{1}{2} \log \mathbf{u}^\top D_{\boldsymbol{\alpha}} Q D_{\boldsymbol{\alpha}} \log \mathbf{u} + \boldsymbol{l}^\top D_{\boldsymbol{\alpha}} \log \mathbf{u}\right\} C_{\mathbf{a}}(\boldsymbol{\alpha}, Q, \boldsymbol{l})} f_{\mathbf{ua}}(\mathbf{z}; \boldsymbol{\alpha}, Q, \boldsymbol{l} + QD_{\boldsymbol{\alpha}} \log \mathbf{u}) \end{aligned}$$

whence statement (i) follows. The change of variable $\tilde{\mathbf{z}} = \mathbf{z}^{\boldsymbol{\beta}}$ implies

$$\mathbb{P}(\mathbf{Z}^{\boldsymbol{\beta}} \in A) = \int_A f_{\mathbf{a}}(\tilde{\mathbf{z}}^{1/\boldsymbol{\beta}}; \boldsymbol{\alpha}, Q, \boldsymbol{l}) \prod_{i=1}^d \beta_i^{-1} \tilde{z}_i^{1/\beta_i - 1} \mathrm{d}\tilde{\mathbf{z}}$$

and simple computations result in

$$f_{\mathbf{a}}(\mathbf{z}^{1/\boldsymbol{\beta}}; \boldsymbol{\alpha}, Q, \boldsymbol{l}) \prod_{i=1}^d \beta_i^{-1} z_i^{1/\beta_i - 1} = \frac{C_{\mathbf{a}^{\boldsymbol{\beta}}}(\boldsymbol{\alpha}/\boldsymbol{\beta}, Q, \boldsymbol{l})}{C_{\mathbf{a}}(\boldsymbol{\alpha}, Q, \boldsymbol{l}) \prod_{i=1}^d \beta_i} f_{\mathbf{a}^{\boldsymbol{\beta}}}(\mathbf{z}; \boldsymbol{\alpha}/\beta, Q, \boldsymbol{l})$$

whence statement (ii) follows. □

We deduce a simple relation between generalised HR-Pareto distribution and (standard) HR-Pareto distribution.

Corollary 3.2. *Let* $\mathbf{Z} \rightsquigarrow \mathrm{HRPar}_{\mathbf{a}}(\boldsymbol{\alpha}, Q, \boldsymbol{l})$ *with* $\boldsymbol{\alpha} \in (0,\infty)^d$. *We have* $\mathbf{Z} \stackrel{d}{=} \mathbf{a}\tilde{\mathbf{Z}}^{c/\boldsymbol{\alpha}}$ *where* $\tilde{\mathbf{Z}} \rightsquigarrow \mathrm{HRPar}_{\mathbf{1_d}}(Q, \boldsymbol{l} - QD_{\boldsymbol{\alpha}} \log \mathbf{a})$ *with exponent* $c > 0$. *Moreover, we have the relationships*

(i) $C_{\mathbf{ua}}(\boldsymbol{\alpha}, Q, \boldsymbol{l} + Q \log \mathbf{u}) = \exp\left\{\frac{1}{2} \log \mathbf{u}^\top D_{\boldsymbol{\alpha}} Q D_{\boldsymbol{\alpha}} \log \mathbf{u} + \boldsymbol{l}^\top D_{\boldsymbol{\alpha}} \log \mathbf{u}\right\} C_{\mathbf{a}}(\boldsymbol{\alpha}, Q, \boldsymbol{l})$

(ii) $C_{\mathbf{a}^{\boldsymbol{\beta}}}(\boldsymbol{\alpha}/\boldsymbol{\beta}, Q, \boldsymbol{l}) = C_{\mathbf{a}}(\boldsymbol{\alpha}, Q, \boldsymbol{l}) \prod_{i=1}^d \beta_i$.

The following proposition relates the moments of generalised Hüsler-Reiss Pareto model with those of the Hüsler-Reiss Pareto model.

Proposition 3.5. *Without loss of generality, assume* $\mathbf{a} = \mathbf{1_d}$ *and let* $\mathbf{Z} \rightsquigarrow \mathrm{HRPar}(\boldsymbol{\alpha}, Q, \boldsymbol{l})$. *Then, the expectation and the covariance matrix of* $\log Z$ *are given by*

$$\mathbb{E}_{\boldsymbol{\alpha}, Q, \boldsymbol{l}}[\log Z_i] = \alpha_i^{-1} \mathbb{E}_{\mathbf{1}_d, Q, \boldsymbol{l}}[\log Z_i] \quad i = 1, \cdots, d$$

and

$$\mathrm{Cov}_{\boldsymbol{\alpha},Q,\boldsymbol{l}}\left(\log Z_i, \log Z_j\right) = \alpha_i^{-1}\alpha_j^{-1}\mathrm{Cov}_{\mathbf{1}_d,Q,\boldsymbol{l}}\left(\log Z_i, \log Z_j\right)$$

where $\mathbb{E}_{Q,\boldsymbol{l}}$ *and* $\mathrm{Cov}_{Q,\boldsymbol{l}}$ *are the expectation and covariance of Hüsler-Reiss Pareto distribution with exponent* 1.

Proof. Proposition 3.4 yields $\mathbb{E}_{\boldsymbol{\alpha},Q,\boldsymbol{l}}[\log Z_i] = \int \alpha_i^{-1}\log z_i f_{\mathbf{1_d}}(\mathbf{z};\mathbf{1_d},Q,\boldsymbol{l})\mathrm{d}\mathbf{z}$. Similarly, we have $\mathbb{E}_{\boldsymbol{\alpha},Q,\boldsymbol{l}}[\log Z_i \log Z_j] = \alpha_i^{-1}\alpha_j^{-1}\mathbb{E}_{Q,\boldsymbol{l}}[\log Z_i \log Z_j]$. Thus the result. □

Remark 3.5. *The family of the generalised Hüsler-Reiss Pareto distributions form a curved exponential family with minimal sufficient statistic* T *given by*

$$T(\mathbf{z}) = \left(\log \mathbf{z} \log \mathbf{z}^\top, \log \mathbf{z}\right).$$

The associated natural parameter space contain positive definite matrices and the set of parameters of interest $(\boldsymbol{\alpha}, Q, \boldsymbol{l})$ *is included in the boundary of the natural parameter space, making the theory difficult.*

3.3.2 Maximum likelihood inference

We assume without loss of generality that $\mathbf{a} = \mathbf{1_d}$ is the known threshold. Based on independent observation $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \cdots$ with distribution $\mathrm{HRPar}_{\mathbf{1_d}}(\theta_0)$, $\quad \theta_0 \in \theta$ we define the log-likelihood

$$L_n(\theta; \mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(n)}) = \sum_{i=1}^{n} \log f_{\mathbf{1_d}}(\mathbf{z}^{(i)}, \theta), \quad \theta \in \Theta,$$

and consider maximum likelihood estimation. It should be noted that we were not able to apply directly the 'classical' maximum likelihood estimation theory from Lehman [120] that uses differentiability properties of the likelihood. Indeed, despite some substantial efforts, we could not prove the relations

$$\frac{\partial^k}{\partial\theta^k}\int_{(0,\infty)^d} f_{\mathbf{1_d}}(\mathbf{z};\theta)\mathrm{d}\mathbf{z} = \int_{(0,\infty)^d} \frac{\partial^k}{\partial\theta^k} f_{\mathbf{1_d}}(\theta,\mathbf{z})\mathrm{d}\mathbf{z}, \quad k = 1,2,3,$$

that are required (assumption M7 in [120, Theorem 7.5.2]). Instead, we use differentiability in quadratic mean and local expansion of the likelihood process as in van der Vaart [176, Chapter 5].

Proposition 3.6. *The statistical model* $\{f_{\mathbf{1_d}}(\theta;\mathbf{z}), \theta \in \Theta\}$ *is differentiable in quadratic mean. Furthermore, the local likelihood process defined by*

$$\tilde{L}_n(h) = L_n\left(\theta_0 + h/\sqrt{n}; \mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(n)}\right), \quad \theta_0 + h/\sqrt{n} \in \Theta,$$

satisfies, uniformly on compact sets,

$$\tilde{L}_n(h) = \tilde{L}_n(0) + \frac{\partial \tilde{L}_n}{\partial h}(0)^\top h - \frac{1}{2} h^\top I_{\theta_0} h + o_p(1), \tag{3.9}$$

$$\frac{\partial \tilde{L}_n}{\partial h}(h) = \frac{\partial \tilde{L}_n}{\partial h}(0) - I_{\theta_0} h + o_p(1), \tag{3.10}$$

$$\frac{\partial^2 \tilde{L}_n}{\partial h \partial h^\top}(h) = -I_{\theta_0} + o_p(1) \tag{3.11}$$

with I_{θ_0} the Fisher information matrix at θ_0. Furthermore, in Equations (3.9)-(3.10)*,*

$$\frac{\partial \tilde{L}_n}{\partial h}(0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f_{\mathbf{1_d}}}{\partial \theta}(Z^{(i)}, \theta_0) \rightsquigarrow \mathcal{N}(0, I_{\theta_0}) \tag{3.12}$$

and in Equation (3.11)*, the $o_P(1)$-term is even uniform on $\{\|h\| \leq n^{1/2-\varepsilon}\}$ for all $\varepsilon > 0$.*

Proof. Differentiability in quadratic mean is proved thanks to Lemma 7.6 in van der Vaart [176]. It is easily checked that $\theta \mapsto \sqrt{f_{\mathbf{1_d}}(\mathbf{z}; \theta)}$ is continuously differentiable for every $\mathbf{z}$. Then we need to check that, with $\ell(\theta, \mathbf{z}) = \log f_{\mathbf{1_d}}(\theta, \mathbf{z})$, the matrix

$$I(\theta) = \mathbb{E}_\theta \left[\frac{\partial \ell}{\partial \theta}(\theta, \mathbf{Z}) \frac{\partial \ell}{\partial \theta}(\theta, \mathbf{Z})^\top \right]$$

is well defined and continuous in θ. This follows easily from the fact that the log-likelihood has the specific form

$$\frac{\partial \ell}{\partial \theta}(\theta, \mathbf{Z}) = \langle A(\theta), T(\mathbf{Z}) \rangle + B(\theta)$$

with $A(\theta)$, $B(\theta)$ continuous in θ and $T(\mathbf{Z}) = (\log \mathbf{Z} \log \mathbf{Z}^\top, \log \mathbf{Z})$. Since $T(\mathbf{z})$ has moment of all orders that depend continuously on θ (this is true for the exponential family Hüsler-Reiss-Pareto and hence for the generalised Hüsler-Reiss-Pareto distributions), $I(\theta)$ is well defined and continuous in θ. From [176, Lemma 7.6], we deduce that the model is differentiable in quadratic mean. For further reference, note that by [176, Theorem 7.2], we have

$$\mathbb{E}_\theta \left[\frac{\partial \ell}{\partial \theta}(\theta, \mathbf{Z}) \right] = 0 \ , \quad I(\theta) = \mathbb{E} \left[\frac{\partial \ell}{\partial \theta}(\theta, \mathbf{Z}) \frac{\partial \ell}{\partial \theta}(\theta, \mathbf{Z})^\top \right] \tag{3.13}$$

and Equation (3.9) holds for all fixed h (we don't have uniformity at this point).

We now prove the uniform asymptotic expansion (3.11). The change of variable $\theta = \theta_0 + h/\sqrt{n}$ yields

$$\frac{\partial^2 \tilde{L}_n}{\partial h \partial h^\top}(h) = \frac{1}{n} \frac{\partial^2 L_n}{\partial \theta \partial \theta^\top}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ell}{\partial \theta \partial \theta^\top}(\theta; \mathbf{Z}^{(i)}),$$

so that, by the law of large numbers,

$$\frac{\partial^2 \tilde{L}_n}{\partial h \partial h^\top}(0) \xrightarrow{a.s.} \mathbb{E}_{\theta_0} \left[\frac{\partial^2 \ell}{\partial \theta \partial \theta^\top}(\theta_0; \mathbf{Z}) \right] := -J_{\theta_0}, \quad \text{as } n \to \infty. \tag{3.14}$$

We don't know at this point that $J_{\theta_0} = I_{\theta_0}$, this will be proven in a final step. Thanks to the Taylor-Lagrange formula, the second-order derivative increment

$$\frac{\partial^2 \tilde{L}_n}{\partial h \partial h^\top}(h) - \frac{\partial^2 \tilde{L}_n}{\partial h \partial h^\top}(0)$$

has norm upper bounded, for $\|h\| \leq n^{1/2-\varepsilon}$, by

$$Cn^{1/2-\varepsilon} \max_{\|h\| \leq n^{1/2-\varepsilon}} \left\| \frac{\partial^3 \tilde{L}_n}{\partial h^3}(h) \right\| = Cn^{-1-\varepsilon} \max_{\|\theta-\theta_0\| \leq n^{-\varepsilon}} \left\| \frac{\partial^3 L_n}{\partial \theta^3}(\theta) \right\|.$$

The specific form

$$\ell(\theta, \mathbf{Z}) = -\frac{1}{2}\langle \log \mathbf{z} \log \mathbf{z}^\top, D_{\boldsymbol{\alpha}} Q D_{\boldsymbol{\alpha}} \rangle + \langle \log \mathbf{z}, D_{\boldsymbol{\alpha}} \boldsymbol{l} \rangle - \log C_{\mathbf{1_d}}(\theta)$$

implies that the third order derivative is upper bounded by

$$\left\| \frac{\partial^3 L_n}{\partial \theta^3}(\theta, \mathbf{z}) \right\| \leq C_1 + C_2 \left\| \sum_{i=1}^n \log \mathbf{Z}^{(i)} \log \mathbf{Z}^{(i)^\top} \right\|$$

for some constants $C_1, C_2 > 0$ that does not depend on $\mathbf{z} \in [\mathbf{0}, \mathbf{1_d}]^c$ and θ in a neighbourhood of θ_0. We deduce

$$\left\| \frac{\partial^2 \tilde{L}_n}{\partial h \partial h^\top}(h) - \frac{\partial^2 \tilde{L}_n}{\partial h \partial h^\top}(0) \right\| \leq cn^{-\varepsilon} \left(C_1 + C_2 \left\| \frac{1}{n} \sum_{i=1}^n \log Z^{(i)} \log Z^{(i)^\top} \right\| \right). \tag{3.15}$$

By the law of large number, the sample mean converges almost surely so that the right hand side is $O_P(n^{-\varepsilon}) = o_P(1)$ uniformly in $\|h\| \leq n^{1/2-\varepsilon}$. Equations (3.14) and (3.15) together imply Equation (3.11) with J_{θ_0} instead of I_{θ_0} for the moment. Equations (3.10) and (3.9) with J_{θ_0} instead of I_{θ_0} follows from (3.9) by integration with the $o_P(1)$ term uniform on compact set. We have already noticed that differentiability in quadratic mean implies (3.11) with I_{θ_0}, so that necessarily the two asymptotic expansion must coincide and $J_{\theta_0} = I_{\theta_0}$. This proves Equations (3.9), (3.10) and (3.11) in their final form. Finally, in view of (3.13), the asymptotic normality (3.12) is a direct consequence of the central limit Theorem. □

The asymptotic development of the likelihood process stated in Proposition 3.6 together with the Argmax Theorem (see Appendix) allows us to study the properties of the maximum likelihood estimator (existence, consistency, asymptotic normality). An important argument is that, provided I_{θ_0} is definite positive, the asymptotic expansion of the second order derivative (3.11) implies that the local likelihood process $\tilde{L}_n(h)$ is strictly concave on $\{\|h\| < n^{1/2-\varepsilon}\}$ with high probability. As we will see in the proof below, this entails that with high probability, the likelihood process $L_n(\theta)$ as a unique local maximiser in $\{\|\theta - \theta_0\| < n^{-\varepsilon}\}$ that we define as $\hat{\theta}_n^{mle}$.

Theorem 3.3. *Let $\theta_0 \in \Theta$ with I_{θ_0} definite positive and assume the observations $Z^{(1)}, Z^{(2)}, \ldots$ are independent with distribution* $\mathrm{HRPar}_{\mathbf{a}}(\theta_0)$. *Then, there exists a maximum likelihood estimators $\hat{\theta}_n^{mle}$ that is asymptotically normal and efficient, i.e.,*

$$\sqrt{n}(\hat{\theta}_n^{mle} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_{\theta_0}^{-1}) \quad \text{as } n \to \infty.$$

Proof. The proof relies on Proposition 3.6 and the Argmax theorem (van der Vaart [176] Corollary 5.58). Consider the stochastic processes

$$M_n(h) = \tilde{L}_n(h) - \tilde{L}_n(0), \quad \|h\| \leq n^{1/2-\varepsilon}$$

and

$$M(h) = Gh - \frac{1}{2}h^\top I_{\theta_0} h$$

where G is a centered Gaussian random vector with variance I_{θ_0}. Proposition 3.6 implies the convergence of M_n to M in distribution in $L^\infty(K)$ for all compact K. The limit process M is continuous and has a unique maximiser h given by $\hat{h} = I_{\theta_0}^{-1} G \rightsquigarrow \mathcal{N}(0, I_{\theta_0}^{-1})$. Define the maximiser

$$\hat{h}_n = \operatorname*{argmax}_{\|h\| \leq n^{1/2-\varepsilon}} M_n(h),$$

where the Argmax exists because M_n is continuous on a compact set. The Argmax theorem implies that provided $\hat{h}_n$ is tight, $\hat{h}_n \xrightarrow{d} \hat{h}$ as $n \to \infty$.

We now prove the tightness of the sequence $\hat{h}_n$, $n \geq 1$. For all $\delta > 0$, there exists $R > 0$ such that

$$\mathbb{P}(\|\hat{h}\| \leq R) \geq 1 - \delta.$$

The relation

$$M(h) = M(\hat{h}) - \frac{1}{2}(h - \hat{h}) I_{\theta_0} (h - \hat{h})$$

implies

$$M(\hat{h}) - \max_{\|\hat{h} - h\| \geq 1} M(h) \geq \frac{1}{2}\lambda_{\min}$$

with $\lambda_{\min} > 0$ the smallest eigenvalue of I_{θ_0}. Therefore, with probability at least $1 - \delta$, we have

$$\max_{\|h\| = R+1} M(h) \leq M(\hat{h}) - \frac{1}{2}\lambda_{\min}.$$

The convergence in distribution of M_n to M in $L^\infty(K)$ with $K = \{h : \|h\| \leq R+1\}$ implies, for large n,

$$\max_{\|h\|\leq R} M_n(h) - \max_{\|h\|=R+1} M_n(h) \geq \frac{1}{4}\lambda_{\min} \tag{3.16}$$

with probability at least $1-2\delta$. The convergence (3.11) together with the positive definiteness of I_{θ_0} implies that M_n is strictly concave on $\{\|h\| \leq n^{1/2-\varepsilon}\}$ with probability at least $1-\delta$ for n large. Hence, Equation (3.16) implies that the maximiser $\hat{h}_n$ of M_n belongs to $\{\|h\| \leq R+1\}$. We have proved that for large n, $\mathbb{P}(\|\hat{h}_n\| \leq R+1) \geq 1-3\delta$, establishing the tightness of $\hat{h}_n$.

Finally, on the event $\|\hat{h}_n\| \leq R+1$, $\hat{h}_n$ belongs to the interior of $\{\|h\| \leq n^{1/2-\varepsilon}\}$ and is therefore a local maximiser of $\tilde{L}_n$ such that $\frac{\partial \tilde{L}_n}{\partial h}(\hat{h}_n) = 0$. Then $\hat{\theta}_n^{mle} = \theta_0 + \frac{\hat{h}_n}{\sqrt{n}}$ is a local maximiser of L_n such that $\frac{\partial L_n}{\partial \theta}(\hat{\theta}_n^{mle}) = 0$, that is a maximum likelihood estimator. Asymptotic normality of $\hat{\theta}_n$ is a direct consequence of the convergence of $\hat{h}_n$ to $\hat{h}$ since $\sqrt{n}(\hat{\theta}_n - \theta_0) = \hat{h}_n \xrightarrow{d} \hat{h} \sim \mathcal{N}(0, I_{\theta_0}^{-1})$. □

3.3.3 Optimising the likelihood

We have proved in the previous section that, with high probability, the likelihood function L_n is strictly concave on a neighbourhood of θ_0 of size $n^{-\varepsilon}$, $\varepsilon > 0$. However, there is no reason why it should be globally convex. We discuss here two issues associated with the likelihood optimisation. The first is the initialisation of an optimisation algorithm and will be addressed thanks to a simple moment estimator that is $\sqrt{n}$-consistent and can serve as a starting point of optimisation routines. The second point is how we can take advantage of the biconcavity of the problem: although not globally concave, the log-likelihood is biconcave in the sense that both partial applications $\boldsymbol{\alpha} \mapsto L_n(\boldsymbol{\alpha}, Q, \boldsymbol{l})$ and $(Q, \boldsymbol{l}) \mapsto L_n(\boldsymbol{\alpha}, Q, \boldsymbol{l})$ are concave. In this context, it is natural to consider alternate convex optimisation.

Proposition 3.7. *Let $\theta = (\boldsymbol{\alpha}, Q, \boldsymbol{l}) \in \Theta$ and assume the observations $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)} \cdots$ independent with distribution* $\mathrm{HRPar}_{\mathbf{1_d}}(\theta)$. *For $j = 1, \ldots, d$ define*

$$N_{n,j} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_j^{(i)}>1\}} \quad \textit{and} \quad O_{n,j} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_j^{(i)}>1\}} \log Z_j^{(i)}.$$

Then the estimator $\hat{\theta}_0 = (\hat{\boldsymbol{\alpha}}_{\mathbf{0}}, \hat{Q}_0, \hat{\boldsymbol{l}}_{\mathbf{0}})$ defined by

$$\hat{\boldsymbol{\alpha}}_{\mathbf{0}} = (N_{n,j}/O_{n,j})_{1\leq j\leq d} \quad \textit{and} \quad (\hat{Q}_0, \hat{\boldsymbol{l}}_{\mathbf{0}}) = \operatorname*{argmax}_{Q,\boldsymbol{l}} L_n(\hat{\boldsymbol{\alpha}}_{\mathbf{0}}, Q, \boldsymbol{l})$$

is strongly consistent and asymptotically normal.

Proof. For $j = 1, \ldots, d$, the thresholded marginal $Z_j | Z_j > 1$ are distributed according to a Pareto distribution with parameter α_j , so that $\mathbb{E}_\theta[\log Z_j \mid Z_j > 1] = \alpha_j^{-1}$. Hence, by the law of large numbers

$$\frac{N_{n,j}}{O_{n,j}} \xrightarrow{a.s} \frac{\mathbb{P}_\theta(Z_j > 1)}{\mathbb{E}_\theta(1_{\{Z_j>1\}} \log Z_j)} = \Big(\mathbb{E}_\theta[\log Z_j \mid Z_j > 1]\Big)^{-1} = \alpha_j$$

so that $\hat{\boldsymbol{\alpha}}_0$ is a consistent estimator for $\boldsymbol{\alpha}$.

On the other hand, the vector $\mathbf{Z}^{\boldsymbol{\alpha}}$ has Hüsler-Reiss distribution $\mathrm{HRP}(Q, \boldsymbol{l})$, so that Theorem 3.2 suggests the maximum-likelihood estimator

$$(\hat{Q}, \hat{\boldsymbol{l}}) = \operatorname*{argmax}_{Q,\boldsymbol{l}} L_n(\boldsymbol{\alpha}, Q, \boldsymbol{l}) = \Psi\Big(\overline{T}_n(D_{\boldsymbol{\alpha}} \log \mathbf{Z}^{(1)}, \ldots, D_{\boldsymbol{\alpha}} \log \mathbf{Z}^{(n)})\Big),$$

where $\Psi(\bar{t})$ denotes the unique solution of the score equation $\frac{\partial \log C}{\partial \theta}(Q, \boldsymbol{l}) = \bar{t}$. As a general result for full exponential families (see e.g. Barndorff-Nielsen [14]), Ψ is a diffeomorphism. Since $\boldsymbol{\alpha}$ is unknown and estimated by $\hat{\boldsymbol{\alpha}}_0$, we set rather

$$\hat{\theta}_0 = (\hat{Q}_0, \hat{\boldsymbol{l}}_{\mathbf{0}}) = \Psi\Big(\overline{T}_n(D_{\hat{\boldsymbol{\alpha}}_{\mathbf{0}}} \log \mathbf{Z}^{(1)}, \ldots, D_{\hat{\boldsymbol{\alpha}}_{\mathbf{0}}} \log \mathbf{Z}^{(n)})\Big).$$

Some simple computations show

$$\overline{T}_n(D_{\hat{\boldsymbol{\alpha}}_{\mathbf{0}}} \log \mathbf{Z}^{(1)}, \ldots, D_{\hat{\boldsymbol{\alpha}}_{\mathbf{0}}} \log \mathbf{Z}^{(n)}) = D_{\mathbf{N_n}/\mathbf{O_n}}$$

where $\mathbf{N_n}$, $\mathbf{O_n}$ and $\mathbf{N_n}/\mathbf{O_n}$ denotes the vectors with components $N_{n,j}$, $O_{n,j}$ and $N_{n,j}/O_{n,j}$ respectively, and

$$M_n = \frac{1}{n}\sum_{i=1}^{d} \log \mathbf{Z}^{(i)} \quad \text{and} \quad V_n = \frac{1}{n}\sum_{i=1}^{d} \log \mathbf{Z}^{(i)} (\log \mathbf{Z}^{(i)})^\top.$$

Hence $\hat{\theta}_0$ can be written in the form

$$\hat{\theta}_0 = \Theta(\mathbf{N_n}, \mathbf{O_n}, \mathbf{M_n}, V_n)$$

with differentiable function Θ. The law of large number ensures the almost sure convergence of $\mathbf{N_n}, \mathbf{O_n}, \mathbf{M_n}, V_n$ as $n \to \infty$, whence strong consistency $\hat{\theta}_0 \overset{a.s.}{\to} \theta$ follows. The central limit theorem ensures the asymptotic normality of $(\mathbf{N_n}, \mathbf{O_n}, \mathbf{M_n}, V_n)$, whence the asymptotic normality of $\hat{\theta}_0$ is deduced via the δ-method (van der Vaart [176, Theorem 3.1]). □

Theorem 3.4. *Let $\theta_0 = (\boldsymbol{\alpha_0}, Q_0, \boldsymbol{l}_0) \in \Theta$ and assume the observations $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \cdots$ independent with distributions* $\mathrm{HRPar}(\theta_0)$. *Define $\hat{\theta}_0$ as in Proposition* (3.7) *and*

$$V_n = \left\{\theta \in \Theta : \|\theta - \theta_0\| < n^{1/2-\varepsilon}\right\}.$$

Define $\hat{\theta}_n^{mle}$ as the unique minimiser of the negative log-likelihood on V_n, i.e.,

$$\hat{\theta}_n^{mle} = \operatorname*{argmin}_{\theta \in V_n} \quad -L_n(\theta; \mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(n)}).$$

Consider the alternating minimisation estimators $\hat{\theta}^{(i)} = (\hat{b\alpha}^{(i)}, \hat{Q}^{(i)}, \hat{\boldsymbol{l}}^{(i)})$ defined by the recursive algorithm

$$\begin{cases} \hat{\boldsymbol{\alpha}}^{(i+1)} & = \operatorname{argmin}_{\boldsymbol{\alpha}} \quad -L_n(\boldsymbol{\alpha}, \hat{Q}^{(i)}, \hat{\boldsymbol{l}}^{(i)}; \mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(n)}) \\ (\hat{Q}^{(i+1)}, \hat{\boldsymbol{l}}^{(i+1)}) & = \operatorname{argmin}_{Q,\boldsymbol{l}} \quad -L_n(\hat{\boldsymbol{\alpha}}^{(i+1)}, Q, \boldsymbol{l}; \mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(n)}) \end{cases} \quad \textit{for } i > 0 \tag{3.17}$$

and initialised with $\hat{\theta}^{(0)} = \hat{\theta}_0$. *Then, with high probability, the sequence of estimators* $(\hat{\theta}^{(i)})_{i\geq 0}$ *converges almost surely to* $\hat{\theta}_n^{mle}$, *i.e.,*

$$\mathbb{P}\left(\lim_{i\to\infty} \hat{\theta}^{(i)} = \hat{\theta}_n^{mle}\right) \to 1, \quad \text{as } n \to \infty. \tag{3.18}$$

Proof. The starting point estimator writes

$$\hat{\theta}_0 = \theta_0 + \frac{1}{\sqrt{n}}(\sqrt{n}(\hat{\theta}_0 - \theta_0)).$$

Proposition 3.7 and Prohorov's theorem implies that $\hat{\theta}_0 \in V_n$ with high probability. Assuming the log-likelihood strictly concave on V_n, we show by recurrence that each iterate of the alternating minimisation algorithm belongs to V_n. Define the level set

$$\mathcal{L}_i = \{\theta : -L_n(\theta) \leq -L_n(\hat{\theta}^{(i)}) + \delta\}, \quad i \geq 0,$$

where $\delta > 0$ is such that $\mathcal{L}_i \cap \partial V_n = \varnothing$. By convex optimisation theory, the intersection between $\mathcal{L}_i$ and V_n is a convex set. Let B_1 and B_2 be open balls centered at $\hat{\theta}^{(i)}$ and $(\hat{\alpha}^{(i+1)}, \hat{Q}^{(i)}, \hat{l}^{(i)})$ such that B_1 and B_2 are subset of $\mathcal{L}_i$. The biconvex property of $-L_n$ implies that the convex hull $\mathrm{conv}(B_1, B_2)$ is a subset of $\mathcal{L}_i$. It results that $\mathrm{conv}(B_1, B_2) \subset \mathcal{L}_i \cap V_n$. A similar reasoning concludes that $\hat{\theta}^{(i)} \in V_n$ for all $i \geq 0$ and therefore the alternating minimisation estimators $\hat{\theta}^{(i)}$ converge to the unique minimiser in V_n. □

3.3.4 A likelihood ratio test for $\alpha_1 = \cdots = \alpha_d$

Following the development of generalised Pareto models, a natural question that arises when one is given a i.i.d. sample $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(n)}$ with distribution $\mathrm{HRPar}(\boldsymbol{\alpha}, Q, \boldsymbol{l})$ is whether the Pareto model would be enough to modelise the data. The following theorem provides a likelihood ratio test for testing $\alpha_1 = \cdots = \alpha_d$.

Theorem 3.5. *Let* $\theta_0 = (\boldsymbol{\alpha}, Q, \boldsymbol{l}) \in \Theta$ *with* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$. *Let* $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(n)}$ *be i.i.d. with distribution* $\mathrm{HRPar}(\theta_0)$. *Denote by* $\hat{\theta}_n$ *the maximum likelihood estimator in the Generalised Hüsler-Reiss Pareto model and* $\hat{\theta}_0$ *the maximum likelihood estimation in the Hüsler-Reiss Pareto model and define the likelihood log-ratio by*

$$\Delta_n = L_n(\hat{\theta}_n) - L_n(\hat{\theta}_0).$$

Then, under the null hypothesis $\alpha_1 = \cdots = \alpha_d$, *the distribution of* $2\Delta_n$ *converge to a chi-squared distribution with* $d-1$ *degree of freedom, i.e.,*

$$2(L_n(\hat{\theta}_n) - L_n(\hat{\theta}_0)) \xrightarrow{d} \chi^2(d-1).$$

Proof. Denote by Θ_0 the subset of Θ defined by $\Theta_0 = \{(\boldsymbol{\alpha}, Q, \boldsymbol{l}) \in \Theta : \alpha_1 = \cdots = \alpha_d\}$. Consider the local log-likelihood process $\tilde{L}_n$ and its maximiser $\hat{h}_n$ on Θ. Likewise, denote by $\hat{h}_n^0$

the maximiser of $\tilde{L}_n$ on Θ_0. We prove below that $2(\tilde{L}_n(\hat{h}_n) - \tilde{L}_n(\hat{h}_n^0)) \xrightarrow{d} \chi^2(p-1)$. Simple calculations imply that the Taylor expansion of $\tilde{L}_n$ at $\hat{h}_n$ writes

$$\tilde{L}_n(h) = \tilde{L}_n(\hat{h}_n) - \frac{1}{2}(h - \hat{h}_n) I_{\theta_0}(h - \hat{h}_n) + o_p(1)$$

where the o_p term is uniform on compact sets containing $\hat{h}_n$. Taking a compact K large enough to contain both $\hat{h}_n$ and $\hat{h}_n^0$, we have

$$\begin{aligned} 2\left(\tilde{L}_n(\hat{h}_n) - \tilde{L}_n(\hat{h}_n^0)\right) &= \min_{h \in K \cap \Theta_0} 2\left(\tilde{L}_n(\hat{h}_n) - \tilde{L}_n(h)\right) \\ &= \min_{h \in K \cap \Theta_0} (h - \hat{h}_n) I_{\theta_0}(h - \hat{h}_n) + o_p(1) \end{aligned}$$

Defining $\langle \cdot, \cdot \rangle_{I_{\theta_0}}$ as the inner product induced by I_{θ_0}, i.e., $\langle a, b \rangle_{I_{\theta_0}} = a^\top I_{\theta_0} b$, we get

$$2\left(\tilde{L}_n(\hat{h}_n) - \tilde{L}_n(\hat{h}_n^0)\right) = \min_{h \in K \cap \Theta_0} \|h - \hat{h}_n\|_{I_{\theta_0}}^2 + o_p(1)$$

The minimum is reached for h the orthogonal projection of $\hat{h}_n$ into Θ_0 for the $\|.\|_{I_{\theta_0}}$ norm. Thus, we have

$$2\left(\tilde{L}_n(\hat{h}_n) - \tilde{L}_n(\hat{h}_n^0)\right) = \sum_{i=1}^{p-1} \langle \hat{h}_n, e_i \rangle_{I_{\theta_0}}^2 + o_p(1)$$

where $(e_1, \cdots, e_{p-1})$ is an orthonormal basis of $\Theta_0^\perp$. Theorem (3.3) implies

$$\left(\langle \hat{h}_n, e_i \rangle_{I_{\theta_0}}\right)_{1 \le i \le p-1} \xrightarrow{d} \mathcal{N}(0_{p-1}, I_{p-1})$$

which in turn results in

$$2\left(\tilde{L}_n(\hat{h}_n) - \tilde{L}_n(\hat{h}_n^0)\right) \xrightarrow{d} \chi^2(p-1).$$

□

Appendices

3.A Lemmas

Lemma 3.2. *Let $\mathcal{S}_d$ denote the linear space of symmetric $d \times d$ matrices and P the linear operator defined as*

$$P : \mathcal{S}_d \to \mathcal{S}_d, \quad Q \mapsto (I - \frac{1}{d}\mathbf{1_d}\mathbf{1_d}^\top)Q(I - \frac{1}{d}\mathbf{1_d}\mathbf{1_d}^\top).$$

Then P is the orthogonal projection on the linear subspace $\mathcal{S}_d^0 = \{S \in E : S\mathbf{1_d} = \mathbf{0}\}$.

Proof. Let $S \in E$, we have

$$\begin{aligned}
P^2(S) &= \left(I - \frac{1}{d}\mathbf{1_d}\mathbf{1_d}^\top\right)^2 S \left(I - \frac{1}{d}\mathbf{1_d}\mathbf{1_d}^\top\right)^2 \\
&= \left(I - \frac{2}{d}\mathbf{1_d}\mathbf{1_d}^\top + \frac{1}{d^2}\mathbf{1_d}\mathbf{1_d}^\top\mathbf{1_d}\mathbf{1_d}^\top\right) S \left(I - \frac{2}{d}\mathbf{1_d}\mathbf{1_d}^\top + \frac{1}{d^2}\mathbf{1_d}\mathbf{1_d}^\top\mathbf{1_d}\mathbf{1_d}^\top\right) \\
&= \left(I - \frac{1}{d}\mathbf{1_d}\mathbf{1_d}^\top\right) S \left(I - \frac{1}{d}\mathbf{1_d}\mathbf{1_d}^\top\right) \\
&= P(S).
\end{aligned}$$

Therefore P is idempotent.
For $S \in \mathcal{S}_d^0$, we have

$$\begin{aligned}
P(S) &= \left(I - \frac{1}{d}\mathbf{1_d}\mathbf{1_d}^\top\right) S \left(I - \frac{1}{d}\mathbf{1_d}\mathbf{1_d}^\top\right) \\
&= S - \frac{2}{d}S\mathbf{1_d}\mathbf{1_d}^\top + \frac{1}{d^2}\mathbf{1_d}\mathbf{1_d}^\top S\mathbf{1_d}\mathbf{1_d}^\top \\
&= S.
\end{aligned}$$

Therefore P acts as the identity on $\mathcal{S}_d$.
For $S \in (\mathcal{S}_d^0)^\perp$, we have

$$P(S) = ... = 0$$

Therefore P is null on $(\mathcal{S}_d^0)^\perp$. This concludes the proof. □

3.B Argmax theorem

We recall the Argmax theorem as in [176]

Theorem 3.6. *Let M_n and M be stochastic processes indexed by subsets H_n and H of a given metric space such that, for every pair of a closed set F and a set K in a given collection $\mathcal{K}$,*

$$(M_n(F \cap K \cap H_n), M_n(K \cap H_n)) \rightsquigarrow (M(F \cap K \cap H), M(K \cap H)). \tag{3.19}$$

Furthermore, suppose that every sample path of the process $h \mapsto M(h)$ possesses a well-separated point of maximum $\hat{h}$ in that, for every open set G and every $K \in \mathcal{K}$,

$$M(\hat{h}) > M(G^c \cap K \cap H), \quad \text{if } \hat{h} \in G, a.s.. \tag{3.20}$$

If $M_n(\hat{h}_n) \geq M_n(H_n) - o_P(1)$ and for every $\varepsilon > 0$ there exists $K \in \mathcal{K}$ such that $\sup_n \mathbb{P}(\hat{h}_n \notin K) < \varepsilon$ and $\mathbb{P}(\hat{h} \notin K) < \varepsilon$, then $\hat{h}_n \rightsquigarrow \hat{h}$.

Proof. If $\hat{h}_n \in F \cap K$, then $M_n(F \cap K \cap H_n) \geq M_n(B) - o_p(1)$ for any set B. Hence, for every closed set F and every $K \in \mathcal{K}$,

$$\mathbb{P}(\hat{h}_n \in F \cap K) \leq \mathbb{P}\left(M_n(F \cap K \cap H_n) \geq M_n(K \cap H_n) - o_p(1)\right) \tag{3.21}$$

$$\leq \mathbb{P}\left(M(F \cap K \cap H) \geq M(K \cap H)\right) + o(1), \tag{3.22}$$

by Slutsky's lemma and the portmanteau lemma (on weak convergence). If $\hat{h} \in F^c$, then $M(F \cap K \cap H)$ is strictly smaller than $M(\hat{h})$ by (3.20) and hence on the intersection with the event in the far right side $\hat{h}$ cannot be contained in $K \cap H$. It follows that

$$\limsup \mathbb{P}(\hat{h}_n \in F \cap K) \leq \mathbb{P}(\hat{h} \in F) + \mathbb{P}(\hat{h} \notin K \cap H). \tag{3.23}$$

By assymption we can choose K such that the left and right sides change by less than ε if we replace K by the whole space. Hence $\hat{h}_n \rightsquigarrow \hat{h}$ by the portmanteau lemma. □

And as a corollary, we have

Corollary 3.3. *Suppose that $M_n \rightsquigarrow M$ in $\ell^\infty(K)$ for every compact subset K of $\mathbb{R}^k$, for a limit process M with continuous sample paths that have unique points of maxima $\hat{h}$. If $H_n \to H, M_n(\hat{h}_n) \geq M_n(H_n) - o_P(1)$, and the sequence $\hat{h}_n$ is uniformly tight, then $\hat{h}_n \rightsquigarrow \hat{h}$.*

Chapter 4

Numerical study

4.1 Introduction

In this part, we will illustrate the results obtained in the chapter 3 on the convergence and asymptotic normality of the maximum likelihood estimator in the Hüsler-Reiss Pareto model. We also assess, thanks to a Monte-Carlo study, the finite sample properties of the estimator (bias, variance). Finally, we study the properties of the maximum likelihood estimator when the sample is approximately HRPareto distributed, i.e. the sample is built as in Proposition 2.1. More details on the experimental protocols will be given in the relevant sections. All experiments were reproduced 1000 times to obtain the Monte-Carlo sample.

4.2 Numerical simulation: bias and variance in the exact simulation case

In this section, the sample is exactly simulated under the Hüsler-Reiss Pareto distribution. Our first aim is to illustrate the effect of the dimension and α on the properties of the maximum likelihood estimator. Therefore, in our first experiment, we will take the dimension parameter d from 2 to 5 and the α parameter equal to $0.5, 1$ or 1.2. From these parameter, we set the distribution parameters $(Q, \boldsymbol{l})$ to satisfy $Q = I_d - \mathbf{1_d}\mathbf{1_d}^\top/d$ and $\boldsymbol{l} = -\alpha/d\ 1_d$. Such construction gives a really "symmetric" structure to our distribution. Finally, we will also illustrate the convergence speed by taking the sample size n from 10 to 1000. Since the number of distribution parameters vary with respect to d, to obtain comparable results, we compute the bias and variance of $\hat{\alpha}$ and $\hat{Q}_{11}$. Thus, we obtain the following results

		$\alpha = 0.5$				$\alpha = 1.0$				$\alpha = 1.2$			
		$\hat{\alpha}$		$\hat{Q}_{11}$		$\hat{\alpha}$		$\hat{Q}_{11}$		$\hat{\alpha}$		$\hat{Q}_{11}$	
d=2	n=10	-65	39	-161	121	-120	143	-133	66	-136	195	-126	52
	n=50	-10	5	-25	8	-22	20	-21	6	-37	29	-24	6
	n=100	-7	3	-13	4	-13	10	-8	3	-19	14	-12	3
	n=1000	-1	1	-2	1	-1	1	-1	1	-2	1	-1	1
d=3	n=10	-54	36	-505	560	-123	138	-379	232	-123	138	-379	232
	n=50	-11	5	-140	24	-15	20	-100	16	-15	20	-100	16
	n=100	1	3	-103	11	3	9	-66	7	3	10	-66	7
	n=1000	3	1	-79	1	10	1	-50	1	10	1	-50	1
d=4	n=10	-54	35	-993	1350	-112	133	-697	739	-112	183	620	726
	n=50	-5	5	-238	35	8	16	-170	24	-5	27	-148	21
	n=100	3	3	-188	15	17	8	-122	10	23	12	-104	8
	n=1000	7	1	-149	1	24	1	-91	1	29	1	-74	1
d=5	n=10	-53	46	-1555	4064	-91	138	-1170	3367	-90	157	-1010	1839
	n=50	3	5	-327	66	11	16	-223	43	17	24	-192	40
	n=100	6	2	-255	26	25	8	-163	18	33	11	-149	15
	n=1000	11	2	-201	2	38	1	-127	1	48	1	-103	1

Table 4.1: Bias and variance: figures where multiplied by 1000

On the estimation viewpoint, since the minimisation problem is a convex problem in the Hüsler-Reiss Pareto model, we use the non linear minimisation routine of R. In spite of the fact the problem is convex, we have to remark that numerical instability arise when we initialise the algorithm near the domain boundary. Obviously, bigger sample implies better estimation but the more surprising result comes from the influence of α on the estimation. We observe that larger values of α yields worse estimation for $\boldsymbol{l}$ and better estimation for the matrix Q. Surprisingly, the variance for $\hat{\alpha}$ is quite stable with respect to d. The same cannot be said for $\hat{Q}_{11}$ which properties worsen as d increase. Another remark is that the estimator is Q_{11} has negative bias and, by construction, the off-diagonal component of Q have positive bias.

Our next aim is to study in details the case $d = 2$ where the distribution has only three parameters (Q_{11}, l_1, l_2). The lower amount of parameter can be tracked and , in that case, we can study the properties under an asymmetric structure on $\boldsymbol{l}$, i.e. $l_1 = -\alpha/2 + \varepsilon$ and $l_2 = -\alpha/2 - \varepsilon$. We will set $\alpha = 1$ and $Q_{11} = 1/2$ for this experiment. We then obtain the following result

		$\hat{Q}_{11}$		$\hat{l}_1$		$\hat{l}_2$	
ε=0	n=10	-140	63	-58	94	-53	101
	n=50	-21	6	-5	11	-19	11
	n=100	-12	3	-9	6	-5	6
	n=1000	-1	1	0	1	-1	1
ε=0.1	n=10	-117	56	-22	102	-83	108
	n=50	-20	6	1	12	-15	12
	n=100	-9	2	-1	6	-10	5
	n=1000	-1	1	-1	1	0	1
ε=0.2	n=10	-140	77	8	150	-138	139
	n=50	-24	6	3	14	-23	14
	n=100	-12	3	0	5	-11	6
	n=1000	0	1	0	1	0	1
ε=0.3	n=10	-146	74	56	190	-166	175
	n=50	-24	6	8	17	-24	16
	n=100	-10	3	-1	8	-13	7
	n=1000	-1	1	1	1	-2	1

Table 4.2: Bias and variance in the asymmetric case: figures where multiplied by 1000

As for the second experiment, the asymmetric structure introduce some bias into the estimation of $\boldsymbol{l}$ that is more obvious for smaller sample but this bias quickly decrease as the sample size increase.

4.3 Numerical simulation: bias and variance in the domain of attraction simulation case

In this section, the sample is taken in the domain of attraction of a Hüsler-Reiss Pareto distribution using Proposition 2.1. That is, let $\mathbf{Z} \sim \mathcal{LN}(\mathbf{m}, \Sigma)$ with $\mathbf{m} \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ positive definite and R an α-Pareto-distributed random variable. Then, since R is a non negative regularly varying random variable with index $-\alpha$, by Proposition 2.1, the product $\mathbf{X} = R\mathbf{Z}$ is in the domain of attraction of the Hüsler-Reiss max-stable model. We simulate a sample S of random vectors $\mathbf{X} = R\mathbf{Z}$ and then we take the observations exceeding the sample quantile $q_S(\varepsilon)$ of order $1-\varepsilon$ to obtain our n-sample S_1 after dividing by $q_S(\varepsilon)$. Thus, the n-sample follows approximately a HRPar$(Q, \boldsymbol{l})$ distribution where the approximation is better as ε is smaller. Moreover, the parameters $(\mathbf{m}, \Sigma)$ where taken such that $(Q, \boldsymbol{l}) = (I_d - \mathbf{1_d}\mathbf{1_d}^\top, -\alpha/d\ \mathbf{1_d})$. As in the last section, we can study the effect of the dimension d and the effect of the parameter α for different sample size. We first consider the case where $\varepsilon = 0.01$ and we obtain the following result

		$\alpha = 0.5$				$\alpha = 1$				$\alpha = 1.2$			
		$\hat{\alpha}$		$\hat{Q}_{11}$		$\hat{\alpha}$		$\hat{Q}_{11}$		$\hat{\alpha}$		$\hat{Q}_{11}$	
d=2	$n = 10$	-59	34	-473	242	-101	142	-421	171	-143	219	-437	145
	$n = 50$	-12	5	-256	18	-20	21	-254	14	-33	31	-255	12
	$n = 100$	-2	2	-240	8	-6	10	-239	5	-10	16	-237	5
	$n = 1000$	-1	1	-222	1	-1	1	-221	1	1	1	-222	1
d=3	$n = 10$	-73	43	-890	873	-122	133	-800	853	-143	208	-741	438
	$n = 50$	-13	5	-37	39	-19	20	-359	29	-41	29	-361	26
	$n = 100$	-5	2	-320	15	-13	9	-326	12	-7	12	-318	11
	$n = 1000$	-1	1	-298	1	-2	1	-299	1	0	1	-296	1
d=4	$n = 10$	-52	34	-1183	1627	-141	157	-1268	1815	-145	201	-1062	1235
	$n = 50$	-8	5	-449	55	-20	20	-431	43	-31	30	-411	37
	$n = 100$	-4	2	-379	20	-10	9	-382	17	-5	14	-370	16
	$n = 1000$	-1	1	-338	1	0	1	-336	1	0	1	-336	1
d=5	$n = 10$	-59	39	-1599	4121	-110	146	-1508	2602	–148	226	-1454	2820
	$n = 50$	-13	5	-480	70	-32	21	-468	65	-17	28	-463	58
	$n = 100$	-2	2	-404	33	-6	9	-410	26	-4	14	-397	26
	$n = 1000$	1	1	-349	4	8	2	-351	3	-14	2	-347	3

Table 4.3: Bias and variance when $\varepsilon = 0.01$: figures where multiplied by 1000

In this set-up, we see that the estimators behaves as in the exact simulation case, i.e. worse behaviour for $\hat{\alpha}$ and better behaviour for $\hat{Q}_{11}$ when α increase and stability of $\hat{\alpha}$ with respect to the dimension. Though the result are in general worse than in the exact simulation case. Even though $\hat{\alpha}$ has similar bias and variance as in the exact simulation case, $\hat{Q}_{11}$ has worse bias and moreover this bias barely decrease as the sample size increase.

In our next experiment, we take $\varepsilon = 0.001$ to obtain the following results

		$\alpha = 0.5$				$\alpha = 1$				$\alpha = 1.2$			
		$\hat{\alpha}$		$\hat{Q}_{11}$		$\hat{\alpha}$		$\hat{Q}_{11}$		$\hat{\alpha}$		$\hat{Q}_{11}$	
d=2	$n = 10$	-64	41	-467	218	-103	148	-414	157	-134	203	-415	134
	$n = 50$	-11	5	-261	18	-22	22	-258	14	-25	30	-257	12
	$n = 100$	-5	2	-238	8	-17	10	-242	6	-10	14	-236	6
	$n = 1000$	-1	1	-223	1	-1	1	-222	1	-1	1	-222	1
d=3	$n = 10$	-60	35	-804	648	-124	154	-717	589	-132	184	-692	448
	$n = 50$	10	5	-366	34	-24	21	-356	28	-27	30	-354	27
	$n = 100$	-7	2	-329	15	-12	9	-325	12	-14	14	-317	11
	$n = 1000$	0	1	-297	1	-1	1	-298	1	0	1	-296	1
d=4	$n = 10$	-59	41	-1272	1671	-125	131	-1190	1507	-184	238	-1088	1078
	$n = 50$	-16	5	-435	51	-15	20	-422	39	-28	29	-412	34
	$n = 100$	-6	2	-376	22	-11	9	-382	16	-12	14	-380	15
	$n = 1000$	0	1	-335	1	0	1	-334	-1	0	1	-336	1
d=5	$n = 10$	-68	38	-1813	6228	-139	141	-1708	4147	-161	209	-1586	3899
	$n = 50$	-12	5	-494	86	-25	20	-485	66	-27	28	-441	64
	$n = 100$	-4	3	-414	34	-7	9	-412	26	-20	9	-423	23
	$n = 1000$	0	1	-362	3	-1	1	-361	2	0	1	-358	1

Table 4.4: Bias and variance when $\varepsilon = 0.001$: figures where multiplied by 1000

We see that the result are about the same as in the case where $\varepsilon = 0.01$ which let us conjecture that the convergence to the Hüsler-Reiss Pareto distribution is slow.

Part II

Partie 2

Chapter 5

Feature selection in weakly coherent matrices

5.1 Introduction

In this chapter, all considered matrices will be assumed to have their columns ℓ_2-normalised.

5.1.1 Background on singular value perturbation

Spectrum perturbation after appending a column has been addressed recently in the literature as a key ingredient in the study of graph sparsification [17], control of pinned systems of ODE's [138], the spiked model in statistics [127]; it can also be useful in Compressed Sensing [50] or for the column selection problem [49]. It is also connected to column selection problems in pure mathematics (Grothendieck and Pietsch factorisation and the Bourgain-Tzafriri restricted invertibility problem) [173].

The goal of the present paper is to study this particular perturbation problem in the special context of column subset selection. The column selection problem was proved essential in High Dimensional Data Analysis [122], [188], [22], [118]. [184], etc. Different criteria for column subsect selection have been studied [27]. Deterministic techniques are often preferred over randomised techniques in industrial applications due to repeatability constraints.

5.1.2 Previous approaches to column selection

Several approaches have been extensively discussed in the literature. Other *deterministic* approaches have been studied recently in the pure mathematics literature, namely [160], [187]. However, these approaches are computationally expensive because of the necessity to perform a matrix inversion at each step. The method of [173] combines randomness with semi-definite programming and although very elegant, is not computationally efficient in practice. A quite efficient techniques is the rank-revealing QR decomposition. Table 1 in [28] provides the performance of this approach and compares it with various other methods. Randomised sampling-based approaches sometimes prove to be faster than the deterministic

approaches. For instance, methods based on leverage scores often gives satisfactory results in practice. Note also that CUR decomposition is much related to the Column Selection tasks and the associated methods can be relevant in practice. A very interesting and efficient approach is the simple greedy algorithm presented in [81] and [82]. However, the method of [82] does not allow for control on the smallest singular value of the selected submatrix, a criterion often considered important for selecting sufficiently decorrelated features.

5.1.3 Coherence

The coherence of a matrix X, usually denoted by $\mu(X)$, is defined as

$$\mu(X) = \max_{1 \leq k < l \leq p} |\langle \mathbf{X}_k, \mathbf{X}_l \rangle|$$

with $\mathbf{X}_k$ the k-th column of X. If the coherence is equal to zero, then the matrix is orthogonal. On the other hand, small coherence does not mean that X is close to square and orthogonal. Indeed, as easy computations show, e.g. i.i.d. Gaussian matrices in $\mathbb{R}^{n \times p}$ and with normalised columns can have a coherence of order $\log(p)^{-1}$ even for n of order $\log(p)^3$; see [42, Section 1.1]. Situations where small coherence holds arise often in practice, especially in signal processing [36] and statistics [42]. The coherence of a matrix has attracted renewed interest recently due to its prominent role in Compressed Sensing [40], Matrix Completion [139], Robust PCA [43] and Sparse Estimation in general. The relationship between coherence and how many columns one can extract uniformly at random which build up a robustly invertible submatrix are studied in [52]. When the coherence is not sufficiently small, the results in [52] are not so much useful anymore and we should turn to the problem of extracting one submatrix with largest possible number of columns with smallest possible correlation. Using coherence information in the study of fast column selection procedures is one interesting question to address in this field.

5.1.4 Contribution of the paper

We propose a greedy algorithm for column subset selection and apply this algorithm to some practical problems. Our contribution to the perturbation and the column selection problems focuses on the special setting where the matrix under study has low coherence. Interestingly, standard perturbation results, e.g. [24] do not take into account the potential incoherence of the matrix under study. The results presented in this paper seem to be the first to incorporate such prior information into the analysis of a column subset selection procedure.

Our approach here is based on a new eigenvalue perturbation bound for matrices with small coherence. Previous bounds have been obtained using the famous Gershgorin's circles theorem [10] but Gershgorin's bound is often too crude. Recent advances have been obtained in this direction in [160] and [187].

5.2 Main results

Our main result is a bound on the smallest singular value after appending a column of a given data matrix with potentially small coherence. Our approach is based on a new result about eigenvalue perturbation. Perturbation after appending a column is a special type of perturbation [50]. The goal of the next subsections is to prove refined results of this type for this problem.

Theorem 5.1 is our first main result on perturbation. This result gives a perturbation bound on the spectrum of a submatrix X_{T_0} of a matrix X with T_0 a subset if $\{1, \cdots, p\}$. Corollary 5.1 takes into account the fact that the coherence of a submatrix can be smaller by a factor α than the coherence of the full matrix. This factor α is crucial in the study of greedy algorithms for column selection where at each step, the selected submatrix has better coherence than the full matrix from which it is extracted. Corollary 5.2 proves a bound on the smallest singular value after successively appending several columns. An example where this result will be useful is the application to greedy column selection algorithms where it can provide a relevant stopping criterion.

5.2.1 Appending one vector: perturbation of the smallest non zero eigenvalue

If we consider a subset T_0 of $\{1, \ldots, p\}$ and a submatrix X_{T_0} of X, the problem of studying the eigenvalue perturbations resulting from appending a column $\mathbf{X}_j$ to X_{T_0}, with $j \notin T_0$ can be studied using Cauchy's Interlacing Lemma (see Appendix) as in the following result.

Theorem 5.1. *Let $T_0 \subset \{1, \ldots, p\}$ with $|T_0| = s_0$ and X_{T_0} a submatrix of X. Let $\lambda_1\left(X_{T_0}X_{T_0}^\top\right) \geq ... \geq \lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right)$ be the eigenvalues of $X_{T_0}X_{T_0}^\top$. We have*

$$\lambda_{s_0+1}\left(X_{T_0}X_{T_0}^\top + \mathbf{X}_j\mathbf{X}_j^\top\right) \geq \lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right) - \min\left(\|X_{T_0}^\top\mathbf{X}_j\|_2, \frac{\|X_{T_0}^\top\mathbf{X}_j\|_2^2}{1-\lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right)}\right). \tag{5.1}$$

Proof. Setting $\mathbf{v} = \mathbf{X}_j$

$$A = X_{T_0}X_{T_0}^\top$$

we obtain from Proposition 5.1 that the smallest nonzero eigenvalue of $X_{T_0}X_{T_0}^\top + \mathbf{X}_j\mathbf{X}_j^\top$ is the smallest root of

$$f(x) = 1 - \sum_{i=1}^{n} \frac{\langle \mathbf{v}, \mathbf{u_i}\rangle^2}{x - \lambda_i\left(X_{T_0}X_{T_0}^\top\right)}.$$

We can decompose this function into two terms

$$f(x) = 1 - \sum_{i=1}^{s_0} \frac{\langle \mathbf{v}, \mathbf{u_i}\rangle^2}{x - \lambda_i\left(X_{T_0}X_{T_0}^\top\right)} - \sum_{i=s_0+1}^{n} \frac{\langle \mathbf{v}, \mathbf{u_i}\rangle^2}{x - \lambda_i\left(X_{T_0}X_{T_0}^\top\right)}.$$

Since $\lambda_i\left(X_{T_0}X_{T_0}^\top\right)=0$ for $i=s_0+1,\dots,n$, we get

$$f(x)=1+\sum_{i=1}^{s_0}\frac{\langle\mathbf{v},\mathbf{u_i}\rangle^2}{\lambda_i\left(X_{T_0}X_{T_0}^\top\right)-x}-\sum_{i=s_0+1}^{n}\frac{\langle\mathbf{v},\mathbf{u_i}\rangle^2}{x}.$$

Notice that

$$\sum_{i=1}^{s_0}\langle\mathbf{v},\mathbf{u_i}\rangle^2\leq\frac{1}{\lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right)}\sum_{i=1}^{s_0}\lambda_i\left(X_{T_0}X_{T_0}^\top\right)\ \langle\mathbf{v},\mathbf{u_i}\rangle^2=\frac{1}{\lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right)}\|X_{T_0}^\top\mathbf{v}\|_2^2.$$

Therefore, upper-bounding $\sum_{i=1}^{s_0}\langle\mathbf{v},\mathbf{u_i}\rangle^2$ by $\frac{1}{\lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right)}\|X_{T_0}^\top\mathbf{v}\|_2^2$ and lower-bounding $\sum_{i=s_0+1}^{n}\langle\mathbf{v},\mathbf{u}_i\rangle^2$ by $1-\frac{1}{\lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right)}\|X_{T_0}^\top\mathbf{v}\|_2^2$, we obtain an upper-bound for f on $]0,\lambda_{s_0}(X_{T_0}X_{T_0}^\top[$. Since f is increasing on the set $]0,\lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right)[$, the smallest root of f is larger than the smallest positive root of $\tilde{f}$ with

$$\tilde{f}(x)=1+\ \frac{\|X_{T_0}^\top\mathbf{v}\|_2^2}{\lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right)\left(\lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right)-x\right)}-\frac{1-\lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right)^{-1}\|X_{T_0}^\top\mathbf{v}\|_2^2}{x}.$$

Thus, after some easy calculations, we find that the smallest root of $\tilde{f}$ is the smallest root of $g(x)=-x^2+x(1+\lambda_{s_0}(X_{T_0}{X_{T_0}}^\top))-\lambda_{s_0}(X_{T_0}{X_{T_0}}^\top))+\|X_{T_0}^\top\mathbf{v}\|_2^2$. Hence,

$$\lambda_{s_0+1}\left(X_{T_0}X_{T_0}^\top+\mathbf{v}\mathbf{v}_j^\top\right)\geq\frac{1+\lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right)-\sqrt{(1-\lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right))^2+4\|X_{T_0}^\top\mathbf{v}\|_2^2}}{2}$$

which, using $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ and $\sqrt{1+a}\leq 1+\frac{a}{2}$, easily gives (5.1).

□

This theorem is useful in the case where μ small enough so that $\|X_{T_0}^\top\mathbf{X}_j\|_2^2\leq 1$. In practice, the submatrices X_{T_0} of X have better coherence than X, up to a factor α. Moreover, we have $\|X_{T_0}\mathbf{X}_j\|_2^2\leq s_0\mu^2$. The following corollary rephrases Theorem 5.1 using the parameter α.

Corollary 5.1. *Let X and T_0 be defined as in Theorem 5.1 and assume*

$$\|X_{T_0}^\top\mathbf{X}_j\|_2^2\leq\alpha s_0\mu^2.$$

Then

$$\lambda_{s_0+1}\left(X_{T_0}X_{T_0}^\top+\mathbf{X}_j\mathbf{X}_j^\top\right)\geq\lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right)-\min\left(\sqrt{\alpha s_0\mu^2},\frac{\alpha s_0\mu^2}{1-\lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right)}\right).\qquad(5.2)$$

5.2.2 Successive perturbations

If we append s_1 columns successively to the matrix X_{T_0}, we obtain the following result

Corollary 5.2. *Let $T_0 \subset \{1, \ldots, p\}$ with $|T_0| = s_0$ and X_{T_0} a submatrix of X. Let $T_1 \subset \{1, \ldots, p\}$ with $|T_1| = s_1$ and $T_0 \cap T_1 = \emptyset$. Let*

$$\varepsilon_{min} = \min \left(\sqrt{\alpha\mu^2} \sum_{i=s_0}^{s_0+s_1} \sqrt{i}, \frac{\alpha\mu^2 s_0}{1 - \lambda_{s_0}\left(X_{T_0} X_{T_0}^\top\right)} + \frac{2(1 - \lambda_{s_0}\left(X_{T_0} X_{T_0}^\top\right))}{s_0} \sum_{i=s_0+1}^{s_0+s_1} \frac{i}{i-1} \right). \quad (5.3)$$

Then

$$\lambda_{s_0+s_1}\left(X_{T_0 \cup T_1}^\top X_{T_0 \cup T_1}\right) \geq \lambda_{s_0}\left(X_{T_0} X_{T_0}^\top\right) - \varepsilon_{min} \quad (5.4)$$

5.3 A greedy algorithm for column selection

Greedy algorithm are commonly used for model selection or feature selection, see for example the forward selection algorithm [21] or the forward stagewise selection algorithm [182][79][100]. See also [96] for more references. The analysis in Section 5.2 suggest that a greedy algorithm can be easily devised for efficient column extraction. The idea is quite simple: append the column which minimises the norm of the scalar products with the columns selected up to the current iteration. This algorithm is described with full details in Algorithm 3 below.

Note that Algorithm 3 requires the computation of the smallest eigenvalue at each step, which might be computationally expensive in large dimensional settings.

5.4 Numerical experiments

5.4.1 Extracting representative time series

Time series are ubiquitous in a world where so many phenomena are monitored via sensor networks. One interesting application of greedy column selection is to

- extract representative time series among large datasets and
- understand the intrinsic "dimension" of the dataset, i.e. the maximum number of different dynamics that are present.
- extract potential outliers.

Input: a matrix $X \in \mathbb{R}^{n\times p}$, $\varepsilon > 0$
Output: a submatrix X_T
Set $s = 1$ and choose a random singleton $T = \{j^{(1)}\} \subset \{1, \dots, p\}$.
Set $\eta^{(1)} = 1$.
while $\eta^{(s)} \geq 1 - \varepsilon$ **do**

Set

$$j^{(s)} \in \operatorname{argmin}_{j\in\{1,\dots,p\}\setminus T} \quad \|X_T^\top \mathbf{X}_j\|_2.$$

Set

$$\alpha^{(s)} = \|X_T^\top \mathbf{X}_{j^{(s)}}\|_2^2/(s\mu(X)^2).$$

Set $T = T \cup \{j^{(s)}\}$.

Set

$$\eta^{(s+1)} = \eta^{(s)} - \min\left(\sqrt{\alpha^{(s)}\ s}\mu, \frac{\alpha^{(s)}\mu(X)^2 s}{1 - \lambda_s(X_T^\top X_T)}\right).$$

Set $s \leftarrow s + 1$.

end
return X_T.

Algorithm 3: Greedy column selection

In this experiment, we considered a set of 1479 times series of length 39 which consist in non-linear transformation of satellite InSAR data [1]. Then, starting from a random time series, we extracted 150 times series sequentially minimising $\|X_T^\top \mathbf{X}_j\|_2, j \notin T$ at each step. Figure 5.1 shows the behaviour of our algorithm over time. For large μ, we see that the bound provided by Corollary 5.2 are worse than the Gershgorin bound and successive applications of Theorem 5.1 provides again a better bound.

[1] a non-linear transformation was performed in order to make the time-series locations and sources impossible to identify

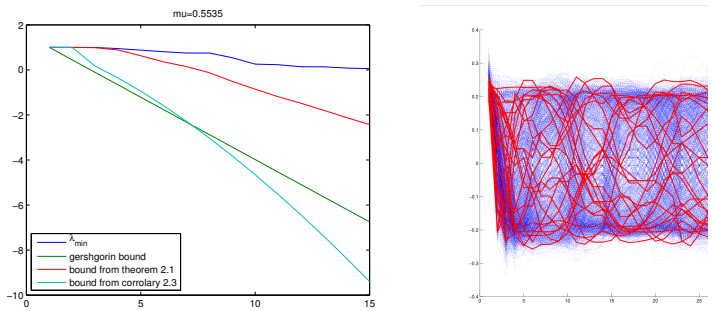


Figure 5.1: Left: Evolution of the smallest singular value in the greedy column selection Algorithm 1. Right: Main extracted Features.

5.4.2 Extracting representative images from a dataset

Extracting representative objects in a dataset is of great importance in data analytics. It can be used to detect outliers or clusters. In this example, we applied our technique to the Yale Faces database shown in Figure 5.2 (Left). In order to cluster the set of images, we performed a preliminary scattering transform [125], [31] of the images in the dataset. We then reshaped the resulting scattering transform matrices into column vectors that we further concatenated into a single matrix X. We selected 9 faces using our column selection algorithm and we obtained the result shown in Figure 5.2 (Right). The total time for this computation was .07 seconds. Larger Pictures are given in the associated report [54].

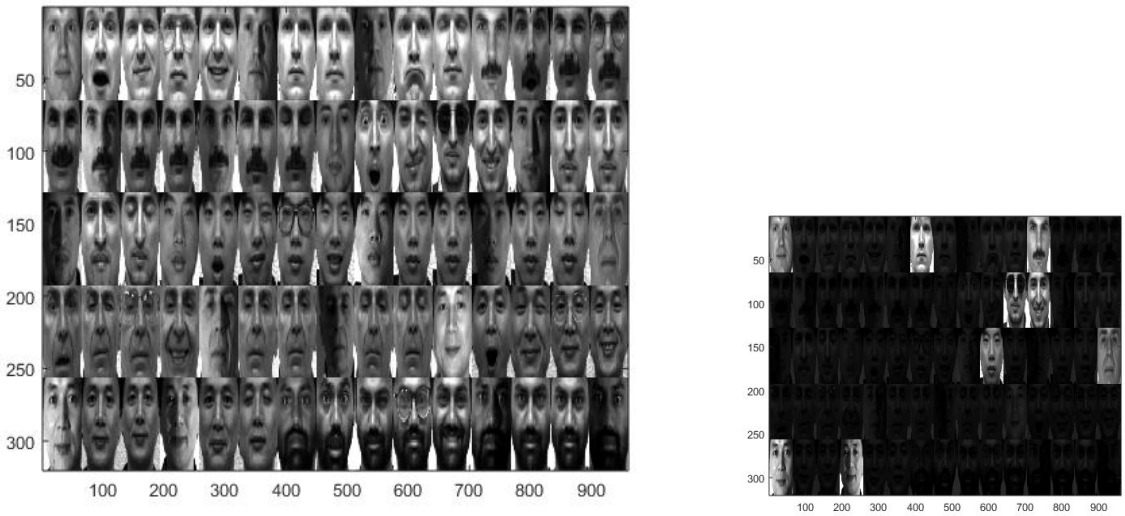


Figure 5.2: Left: Faces from the Yale database. Right: Faces selected by our algorithm.

5.4.3 Comparison with CUR

We compared the behaviour of our method with the CUR algorithm proposed in [28]. We generated 100 matrices with i.i.d. standard Gaussian entries, with 100 rows and 10000 columns and performed both Algorithm 1 from the present paper and the CUR method. We restricted the study to the case of 10 columns to be extracted. The following histograms in Figure 5.3 show the relative performance of our method as compared to CUR [28] [2].

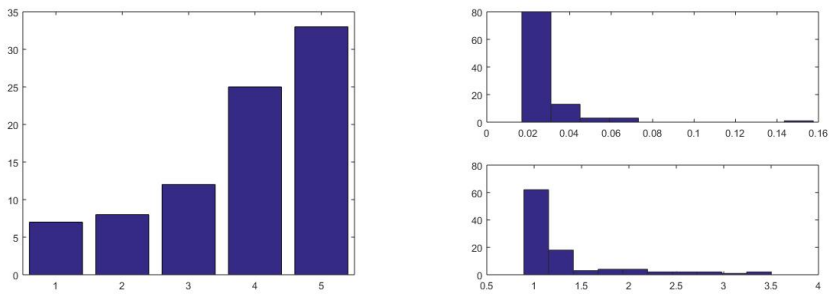


Figure 5.3: Left: counts of the number of singular values of the submatrix extracted with Algorithm 1 larger than for CUR among the 5 smallest singular spectrum for 100 independent Monte Carlo trials. Right-top: histogram of the computation time for Algorithm 1. Right-bottom: histogram of the computation time for the CUR method [28].

The Monte Carlo experiments shown in Figure 5.3 suggest that our method performs better than the CUR method, both from the viewpoint of providing submatrices with larger singular values on average and for a much smaller computational effort (our method was around 50 times faster for these experiments). These experiments are extracted from a more extensive set of experiments, including comparison with other methods, proposed in [54].

5.5 Conclusion and perspectives

In this paper, we established a relationship between the coherence and a perturbation bound for incoherent matrices. Our approach is based on perturbation theory and no randomness assumption on the design matrix is used to establish this property. Coherence plays an important role in many pure and applied mathematical problems and perturbation results may help go significantly further. Two such problems for which we are planning further investigations are the following.

- **Random submatrices are well conditioned**. Matrices with small coherence have a very nice property: most submatrices with s columns have their eigenvalues concentrated around 1 for s of the order $n/\log(p)$. This was first studied in [174], [42, Theorem 3.2 and following comments] and then improved in [52]. The study of such properties is

[2] we used the Matlab implementation provided on Christos Boutsidis webpage

of tremendous importance in the study of designs for sparse recovery [42]. An interesting potential application of studying spectrum perturbations after appending a column is the one of spectrum concentration via the bounded difference inequality [25]. Such concentration bounds should also appear essential in understanding the behaviour of random column sampling algorithms [66], [27].

- **The restricted invertibility problem**. Given any matrix X, the Restricted Invertibility problem of Bourgain and Tzafriri is the one of extracting the largest number of columns X_j, $j \in T$ form X while ensuring that the smallest singular value of X_T stays away from zero. Different procedures have been proposed for this problem. Some of them are randomised and some are deterministic. The original results obtained by Bourgain and Tzafriri were based on random selection [26]. The current best results were recently obtained by Youssef in [187] based on an remarkable inequality discovered by Batson, Spielman and Srivastava in [16]. In [49], using an elementary perturbation approach, S. Chrétien and S. Darses recently obtained a very short proof of a weaker version of the Bourgain-Tzafriri theorem (up to a $\log(s)$ multiplicative term). Our next goal is to refine these types of perturbation results in the small coherence setting and extend the applicability to Big Data analytics.

Appendices

5.A Interlacing and the characteristic polynomial

Recall that for a matrix A in $\mathbb{R}^{n\times n}$, p_A denotes the characteristic polynomial of A.

Proposition 5.1. Cauchy's Interlacing theorem. *If $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and associated eigenvectors $v_1,\ldots,v_n$, and $v \in \mathbb{R}^n$, then*

$$p_{A+\mathbf{v}\mathbf{v}^\top}(x) = p_A(x)\left(1 - \sum_{i=1}^{n} \frac{\langle \mathbf{v}, \mathbf{u_i}\rangle^2}{x-\lambda_i}\right). \tag{5.5}$$

The previous lemma states in particular that the eigenvalues of A interlace those of $A + \mathbf{v}\mathbf{v}^\top$. See [111] for a short proof and other references.

5.B Proof of Corollary 5.2

Define $\lambda_{s_0+s,\min}$ by

$$\begin{cases} \lambda_{s_0,\min} = \lambda_{s_0}\left(X_{T_0}X_{T_0}^\top\right) \\ \lambda_{s_0+s+1,\min} = \lambda_{s_0+s}\left(X_{T_0\cup T}X_{T_0\cup T}^\top\right) - \min\left(\sqrt{\alpha\mu^2(s_0+s)}, \frac{\alpha\mu^2(s_0+s)}{1-\lambda_{s_0+s,\min}}\right) \end{cases}$$

There are two step to prove for the theorem. The first step set up the basis for some recursive relation. We show that, for $s \geq 0$, to obtain a lower-bound of λ_{s_0+s+1}, it is enough to use $\lambda_{s_0+s,\min}$ as the basis for Corollary 5.1. Or simply that we have

$$\begin{aligned} &\lambda_{s_0+s,\min} - \min\left(\sqrt{\alpha\mu^2(s_0+s)}, \frac{\alpha(s_0+s)\mu^2}{1-\lambda_{s_0+s,\min}}\right) \\ &\quad \leq \lambda_{s_0+s}\left(X_{T_0\cup T}X_{T_0\cup T}^\top\right) - \min\left(\sqrt{\alpha\mu^2(s_0+s)}, \frac{\alpha(s_0+s)\mu^2}{1-\lambda_{s_0+s}\left(X_{T_0\cup T}X_{T_0\cup T}^\top\right)}\right) \\ &\quad \leq \lambda_{s_0+s+1}\left(X_{T_{s+1}}X_{T_{s+1}}^\top\right). \end{aligned}$$

It is obvious that the case where one minimum is equal to $\sqrt{\alpha\mu(s_0+s)}$ satisfy the property. Therefore, we study the following inequality

$$\lambda_{s_0+s,\min} - \frac{\alpha(s_0+s)\mu^2}{1-\lambda_{s_0+s,\min}} \leq \lambda_{s_0+s} - \frac{\alpha(s_0+s)\mu^2}{1-\lambda_{s_0+s}\left(X_{T_0\cup T}X_{T_0\cup T}^\top\right)}.$$

It is easily verified that the property is true for $s=0$. Denote

$$\varepsilon = \lambda_{s_0+s}\left(X_{T_0\cup T}X_{T_0\cup T}^\top\right) - \lambda_{s_0+s+1}\left(X_{T_{s+1}}X_{T_{s+1}}^\top\right). \tag{5.6}$$

Then the recursion step is equivalent to proving that

$$\alpha\mu^2\frac{s_0+s}{1-\lambda_{s_0+s}\left(X_{T_0\cup T}X_{T_0\cup T}^\top\right)} + \alpha\mu^2\frac{s_0+s+1}{1-\lambda_{s_0+s+1,\min}} \geq \varepsilon + \alpha\mu^2\frac{s_0+s+1}{1-\lambda_{s_0+s}\left(X_{T_0}X_{T_0}^\top\right)+\varepsilon}. \tag{5.7}$$

This inequality can be interpreted as the sum of errors obtained by applying Corollary 5.1 twice is greater than the sum of errors obtained if we knew the true value after one perturbation then apply Corollary 5.1.

Let g be defined by

$$g_{s_0+s}(x) = x + \alpha\mu^2\frac{s_0+s+1}{1-\lambda_{s_0+s}(X_{T_0\cup T}X_{T_0\cup T}^\top)+x}.$$

Since $\varepsilon \leq \alpha\mu^2(s_0+s)/(1-\lambda_{s_0+s}(X_{T_0\cup T}X_{T_0\cup T}^\top))$ by Corollary (5.1), it is enough to prove g increasing.

A simple analysis show that g is strictly increasing if

$$\alpha\mu^2\frac{s_0+s+1}{(1-\lambda_{s_0+s}(X_{T_0\cup T}X_{T_0\cup T}^\top))^2} < \frac{3}{4}.$$

In the case $\alpha\mu^2(s_0+s+1)((1-\lambda_{s_0+s}(X_{T_0\cup T}X_{T_0\cup T}^\top))^2) > 3/4$, we can show that the left side of Inequation (5.7) is larger than $1-\lambda_{s_0+s}(T_0\cup T)$ and this means that we obtain the trivial bound 0 and therefore of not relevant interest.

For the second part, we aim at bounding the sum of errors. We have

$$\sum_{i=s_0}^{s_0+s}\min\left(\sqrt{\alpha\mu^2 i}, \frac{\alpha\mu^2 i}{1-\lambda_{i,\min}}\right) \leq \min\left(\sum_{i=s_0}^{s_0+s}\sqrt{\alpha\mu^2 i}, \sum_{i=s_0}^{s_0+s}\frac{\alpha\mu^2 i}{1-\lambda_{i,\min}}\right).$$

The second sum writes

$$\sum_{i=s_0}^{s_0+s} \frac{\alpha\mu^2 i}{1-\lambda_{s_0}(X_{T_0}X_{T_0}^\top)+\sum_{j=s_0}^{i-1}\frac{\alpha\mu^2 j}{1-\lambda_{s_0}(X_{T_0}X_{T_0}^\top)}}$$
$$= \sum_{i=s_0}^{s_0+s} \frac{\alpha\mu^2 i}{1-\lambda_{s_0}(X_{T_0}X_{T_0}^\top)+\frac{\alpha\mu^2}{1-\lambda_{s_0}(X_{T_0}X_{T_0}^\top)}\sum_{j=s_0}^{i-1} j}$$

This is equal to

$$\sum_{i=s_0}^{s_0+s} \frac{\alpha\mu^2 i}{1-\lambda_{s_0}(X_{T_0}X_{T_0}^\top)+\sum_{j=s_0}^{i-1}\frac{\alpha\mu^2 j}{1-\lambda_{s_0}(X_{T_0}X_{T_0}^\top)}}$$
$$= \sum_{i=s_0+1}^{s_0+s} \frac{\alpha\mu^2 i}{1-\lambda_{s_0}(X_{T_0}X_{T_0}^\top)+\frac{\alpha\mu^2 s_0(i-1)}{1-\lambda_{s_0}(X_{T_0}X_{T_0}^\top)}} + \frac{\alpha\mu^2 s_0}{1-\lambda_{s_0}(X_{T_0}X_{T_0}^\top)}$$

Simple computations lead to the result.

Therefore applying s_1 times Corollary 5.1 and each time upper-bounding, we have (5.4).

Chapter 6

Small coherence implies the weak Null Space Property

6.1 Introduction

6.1.1 Motivation

Compressed Sensing is a new paradigm for data acquisition which was discovered in [37] and [72] and has had a paramount impact on modern Signal Processing, Statistics, Applied Harmonic Analysis, Machine Learning, to name just a few. The whole field started after it was discovered that if $\boldsymbol{\beta}$ is sufficiently sparse, one could recover the support and sign pattern of a high dimensional vector $\boldsymbol{\beta} \in \mathbb{R}^p$ from just a few linear measurements

$$\mathbf{y} = X\beta + \boldsymbol{\epsilon},$$

where $X \in \mathbb{R}^{n \times p}$, with $n \ll p$, by solving a simple convex programming problem of the form

$$\min_{\mathbf{b} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - X\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_1.$$

In the remainder of this chapter, we will assume that the columns of X are ℓ_2 normalised.

One condition implying that both support and sign pattern can be recovered is called the Restricted Isometry Property (RIP) [35]. More precisely, RIP is the property that for all index subset $T_0 \subset \{1, \ldots, p\}$ with $|T_0| = s_0$, all the singular values of the submatrix X_{T_0} whose columns are the columns of X indexed by T_0, lie in the interval $(1 - \delta, 1 + \delta)$.

One key result relating RIP and recovery of the basic features of a sparse vector is the fact that RIP implies the so-called Null Space Property, which says that the kernel of X

does not contain any sparse vector. More precisely, the NSP is the property that for all $T_0 \subset \{1, \ldots, p\}$ with $|T_0| = s_0$, and for all $\mathbf{h} \in \mathrm{Ker}(X)$,

$$\|\mathbf{h}_{T_0}\|_2 \leq C \; \|\mathbf{h}_{T_0^c}\|_1 / \sqrt{s_0} \tag{6.1}$$

with $C \in (0, 1)$. It is well known that the NSP is the key property behind sparse recovery using Basis Pursuit type of methods, whereas RIP is not. The main reason for introducing the RIP is that it provides a pedagogical step for proving the NSP in the case of random matrices. See [33] for a set of very interesting results in this direction. It was recently shown that the NSP can also be derived without the RIP for random design [6]. Thus, understanding more precisely what are the conditions on the design matrix for which we can obtain a kind of NSP is quite an important question in this field.

Some very interesting work has been published recently in order to test if the NSP or weaker version of this property hold for a given matrix using convex programming; see e.g. [60]. On the other hand, one of the main drawbacks of the Restricted Isometry Property is that one cannot in general check if a given matrix X satisfies it in polynomial time. Therefore, RIP is usually not considered of practical interest. Another property often used in many sparse recovery problems is the property of small coherence.

The coherence of a matrix is an important quantity in the study of designs for sparse recovery is the coherence. It will be denoted by μ, will be defined as

$$\mu = \max_{1 \leq k < l \leq p} |\langle \mathbf{X_k}, \mathbf{X_l} \rangle|. \tag{6.2}$$

If the columns are almost orthogonal, then, one usually expects that the performance of Basis Pursuit should be almost as good as in the orthogonal case. This have been rigorously studied in e.g. [42]. The main motivation for using the coherence is that it is conceptually intuitive and also very easy to compute.

On the other hand, it was also proved in [174], [42, Theorem 3.2 and following comments] that if a matrix X has small coherence, then for most index subsets T_0 with cardinal $|T_0| = s_0$, the singular values of X_{T_0} lie in the interval $(1 - \delta, 1 + \delta)$ [1]. In other words, small coherence implies a kind of weak RIP where the singular value concentration property holds for most instead of all submatrices with s_0 columns from X. However, such results, although conceptually very interesting do not address the main problem of proving NSP type properties.

6.1.2 Goal of the paper

Our aim in the present paper is to understand better the role of the coherence for Compressed Sensing by understanding how a small coherence implies a weaker version of the Null Space Property. The main result of the present work is the following. We prove that if a matrix X has small coherence, then, for most index subsets $T_0 \subset \{1, \ldots, p\}$ with cardinal $|T_0| = s_0$, and for all $\mathbf{h} \in \mathrm{Ker}(X)$, (6.1) holds for some positive C_μ. In other words, small coherence implies a kind of weak Null Space Property which holds for most, instead of all, T_0 with $|T_0| = s_0$.

[1] the precise result underpinning this statement will be recalled in Section 6.2.3 below

6.1.3 Additional notation

For $T \subset \{1, \dots, p\}$, we denote by $|T|$ the cardinal of T. Given a vector $\mathbf{x} \in \mathbb{R}^p$, we set $\mathbf{x_T} = (x_j)_{j \in T} \in \mathbb{R}^{|T|}$. The canonical scalar product in $\mathbb{R}^p$ is denoted by $\langle \cdot, \cdot \rangle$.

For any matrix $A \in \mathbb{R}^{d_1 \times d_2}$, we denote by A^t its transpose. The set of symmetric real matrices is denoted by $\mathbb{S}_n$. We denote by $\|A\|$ the operator norm of A. We use the Loewner ordering on symmetric real matrices: if $A \in \mathbb{S}_n$, $0 \preceq A$ denotes positive semi-definiteness of A, and $A \preceq B$ stands for $0 \preceq B - A$. The singular values of A will be denoted by $\sigma_{\max}(A) = \sigma_1(A) \geq \cdots \geq \sigma_{\min\{d_1,d_2\}} = \sigma_{\min}(A)$.

6.2 Background

In this section, we recall some well known previous results relating coherence, singular value concentration, RIP and NSP. We begin with some definitions.

6.2.1 Weak NSP and weak RIP

Weak Null Space Property

First, the weak-Null Space Property.

Definition 6.1. *A matrix $X \in \mathbb{R}^{n \times p}$ satisfies the Weak Null Space Property weak-NSP(s_0,C,π) if for at least a proportion π of all index subsets $T_0 \subset \{1, \dots, p\}$ with $|T_0| = s_0$, and for all $\mathbf{h} \in \mathrm{Ker}(X)$,*

$$\|\mathbf{h_{T_0}}\|_2 \leq C \ \|\mathbf{h_{T_0^c}}\|_1 / \sqrt{s_0}. \tag{6.3}$$

Notice that when $\pi = 1$, we recover the definition of the standard Restricted Isometry Property.

The main consequence of the weak Null Space Property is that exact recovery holds for the basis pursuit problem. Since the work [56], this can be proved swiftly as follows. Let us first recall the framework: we assume that $\mathbf{y} = X\boldsymbol{\beta}$, i.e. we are in the noise free setting and $\boldsymbol{\beta}$ has support T_0 with $|T_0| \leq s_0$. Then, we solve

$$\min_{\mathbf{b} \in \mathbb{R}^p} \quad \|\mathbf{b}\|_1 \text{ s.t. } \mathbf{y} = X\mathbf{b}.$$

Let $\hat{\boldsymbol{\beta}}$ denote a minimiser. Then, we have

$$\|\hat{\boldsymbol{\beta}}\|_1 \leq \|\boldsymbol{\beta}\|_1,$$

which gives

$$\|\hat{\boldsymbol{\beta}}_{T_0^c} - \boldsymbol{\beta}_{\boldsymbol{T_0^c}}\|_1 \leq \|\hat{\boldsymbol{\beta}}_{T_0} - \boldsymbol{\beta}_{T_0}\|_1 + 2\|\boldsymbol{\beta}_{T_0^c}\| \tag{6.4}$$

and thus, by the Cauchy-Schwartz inequality

$$\|\hat{\boldsymbol{\beta}}_{T_0^c} - \boldsymbol{\beta}_{\boldsymbol{T_0^c}}\|_1 \leq \sqrt{s_0}\ \|\hat{\boldsymbol{\beta}}_{T_0} - \boldsymbol{\beta}_{\boldsymbol{T_0}}\|_2 + 2\|\boldsymbol{\beta}_{\boldsymbol{T_0^c}}\| \tag{6.5}$$

Since $\boldsymbol{\beta}$ has support T_0, we obtain that $\boldsymbol{\beta}_{T_0^c} = 0$. Using the fact that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ lies in the kernel of X and using (6.3), we obtain from (6.5) that $\|\hat{\boldsymbol{\beta}}_{T_0^c} - \boldsymbol{\beta}_{T_0^c}\|_1 = 0$. Using (6.3) again, we conclude that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 = 0$, i.e. exact recovery holds. More results of this type can be found in [35] and [80].

Weak Restricted Isometry Property

The weak-Restricted Isometry Property is the subject of the next definition.

Definition 6.2. *A matrix $X \in \mathbb{R}^{n\times p}$ satisfies the Weak Restricted Isometry Property weak-RIP(s,ρ,π) if for at least a proportion π of all index subsets $T_0 \subset \{1,\ldots,n\}$ with $|T_0| = s_0$,*

$$\sqrt{1-\rho} \leq \sigma_{\min}(X_{T_0}) \leq \cdots \leq \sigma_{\max}(X_{T_0}) \leq \sqrt{1+\rho}. \tag{6.6}$$

Notice that when $\pi = 1$, we recover the definition of the standard Restricted Isometry Property.

6.2.2 On the relationship between RIP and NSP

One of the cornerstones of Compressed Sensing is the Null Space Property. It is well known that RIP implies NSP as stated in the next theorem. We will use the standard notations RIP(s_0,ρ) for RIP(s_0,ρ,1) and NSP(s_0,C) for NSP(s_0,C,1).

Theorem 6.1. [35] *Any matrix $X \in \mathbb{R}^{n\times p}$ satisfying RIP($2s_0$,δ) satisfies NSP(s_0,C) with $C \leq \sqrt{2}(1+\delta)/(1-\delta)$.*

6.2.3 On the relationship between the Coherence and weak-RIP

The first result relating small coherence with weak-RIP was established by [42] based on a result about column selection due to Tropp [174]. A refinement of this result is recalled in the next theorem.

Theorem 6.2. Chrétien and Darses [52] *Let $r \in (0,1)$, $\alpha \geq 1$. Let us be given a full rank matrix $X \in \mathbb{R}^{n\times p}$ and a positive integer s_0, such that*

$$\mu \leq \frac{r}{(1+\alpha)\log p} \tag{6.7}$$

$$s_0 \leq \frac{r^2}{(1+\alpha)e^2}\ \frac{p}{\|X\|^2 \log p}. \tag{6.8}$$

Let $T_0 \subset \{1, \ldots, p\}$ be a random support with uniform distribution on index sets satisfying $|T_0| = s_0$. Then the following bound holds:

$$\mathbb{P}\left(\|X_{T_0}^\top X_{T_0} - I\| \geq r\right) \leq \frac{1944}{p^\alpha}. \tag{6.9}$$

This theorem was used in, e.g. [51] for a study of the LASSO when the variance is unknown. It has been also used in remote sensing [106], in the study of Gaussian erasure channels [136], Kaczmarcz type methods for least squares [129], extensions of RIP [13]; see also [87].

6.2.4 The Gershgorin bound

The Gershgorin theorem gives a bound on the operator norm as a function of the coherence. More precisely, as discussed e.g. in [10], for each index subset $T \subset \{1, \ldots, p\}$ with cardinal $|T_0| = s_0$,

$$\|X_{T_0}^\top X_{T_0} - I\| \leq \mu(s_0 - 1). \tag{6.10}$$

Clearly, this result starts being useful when μ is much smaller than s_0. In the application for the LASSO, it is often assumed that this indeed the case as in e.g. [42].

6.3 Main results: small coherence implies weak-NSP

In this section, we state and prove the main result of this paper, namely that small coherence implies weak-NSP. Our main theorem is the following.

Theorem 6.3. *Let $X \in \mathbb{R}^{n \times p}$, $s_0 \leq n$ and $\alpha > 0$. Assume that*

$$s_0 \leq \frac{1}{16(1+\alpha)e^2} \; \frac{p}{\|X\|^2 \log p}. \tag{6.11}$$

Let μ denote the coherence of X. Let

$$\varepsilon_{min} = \frac{\frac{1}{4}\; s_0^3\; \mu^2 + s_0^{3/2}\; \mu}{(3 - 4 s_0 \mu^2)}$$

$$\varepsilon_{max} = 144 s_0^3\; \mu^2 + 72 s_0^{3/2} \mu.$$

Assume that

$$\mu \leq \min\left\{ \frac{1}{\sqrt{288 s_0^{5/2}\left(2 s_0^{3/2} + 1\right)}}, \frac{1}{\sqrt{\frac{3}{2} s_0^4 + 6 s_0^{5/2} + 2 s_0}}, \frac{1}{4(1+\alpha)\log p} \right\}.$$

Then, the matrix X verifies the weak-NSP(s_0,C,π) with $\pi = 1 - 1944/p^\alpha$ and

$$C = \frac{\lambda_1 - \lambda_{s_0} + 3\ s_0\ (\varepsilon_{max} + \varepsilon_{min})}{\lambda_1 - 3\ s_0\ \varepsilon_{min}}.$$

In particular, if

$$\mu \leq \min\left\{\frac{c_0}{s_0^{5/2}}, \frac{1}{4(1+\alpha)\log p}\right\} \tag{6.12}$$

for some positive constant c_0, then the matrix X verifies the weak-NSP(s_0,C,π) with $\pi = 1 - 1944/p^\alpha$ and

$$\varepsilon_{min} = \frac{1}{4}\frac{c_0^2 s_0^{-2}/4 + c_0 s_0^{-1}}{1/2 - c_0^2 s_0^{-4}}$$

$$\varepsilon_{max} = \frac{1}{4}\frac{144 s_0^{-1} c_0^2 + 72 c_0 s_0^{-2}}{\lambda_1 - 1}$$

Then, the matrix X verifies the weak-NSP(s_0,C,π) with $\pi = 1 - 1944/p^\alpha$ and

$$C = \frac{1 + \frac{3}{4}\left(\frac{c_0^2 s_0^{-1}/4 + c_0}{1/2 - c_0^2 s_0^{-4}} + \frac{144 c_0^2 + 72 c_0 s_0^{-1}}{\lambda_1 - 1}\right)}{1 - \frac{3}{4}\frac{c_0^2 s_0^{-1}/4 + c_0}{1/2 - c_0^2 s_0^{-4}}}. \tag{6.13}$$

Proof. Using Theorem 6.2, for

$$\mu \leq \frac{1}{4(1+\alpha)\log p} \tag{6.14}$$

with probability larger that π, an index subset T_0 with cardinality s_0

$$s_0 \leq \frac{1}{16(1+\alpha)e^2}\ \frac{p}{\|X\|^2 \log p}. \tag{6.15}$$

satisfies

$$\frac{5}{4} \geq \lambda_1 \geq \lambda_{s_0} \geq \frac{3}{4}. \tag{6.16}$$

where

$$\lambda_1 := \lambda_1(X_{T_0} X_{T_0}^\top) \tag{6.17}$$

and

$$\lambda_{s_0} := \lambda_{s_0}(X_{T_0} X_{T_0}^\top). \tag{6.18}$$

Let $h \in \mathrm{Ker}(X)$ and let T_0 be a subset of $\{1, \ldots, p\}$ with cardinality $|T_0| = s_0$ verifying (6.16), (6.17) and (6.18). Define

(i) T_1 as the index set of the s_0 largest entries of $h_{T_0^c}$ in absolute value,

(ii) T_2 as the index set of the s_0 largest entries of $h_{(T_0 \cup T_1)^c}$ in absolute value,

(iii) etc ...

Let J denote the number of subsets obtained in this process [2]. Let $T = T_0 \cup T_1$. By (6.29) in Corollary 6.1, we have that

$$(\lambda_{s_0} - 3\ s_0\ \varepsilon_{min})\ \|\mathbf{h}_T\|_2^2 \leq \|X_T \mathbf{h}_T\|_2^2. \tag{6.19}$$

Moreover, since h belongs to the kernel of X,

$$\begin{aligned} \|X_T \mathbf{h}_T\|_2^2 &= \left|\langle X_T \mathbf{h}_T, X\mathbf{h}\rangle - \langle X_T \mathbf{h}_T, X_{T^c} \mathbf{h}_{T^c}\rangle\right|, \\ &= \left|\sum_{j=2,\ldots,J} \langle X_T \mathbf{h}_T, X_{T_j} \mathbf{h}_{T_j}\rangle\right|. \end{aligned}$$

On the other hand, by Lemma 6.5, we have for $j = 2, \ldots, J$,

$$\langle X_T \mathbf{h}_T, X_{T_j} \mathbf{h}_{T_j}\rangle \leq (\lambda_1 + 3\ s_0\ \varepsilon_{max})\ \|\mathbf{h}_T\|_2\ \|\mathbf{h}_{T_j}\|_2.$$

Therefore,

$$\begin{aligned} \|X_T \mathbf{h}_T\|_2^2 &= \left|\sum_{j=2,\ldots,J} \langle X_T \mathbf{h}_T, X_{T_j} \mathbf{h}_{T_j}\rangle\right| \\ &\leq \sum_{j=2,\ldots,J} \left|\langle X_T \mathbf{h}_T, X_{T_j} \mathbf{h}_{T_j}\rangle\right| \\ &\leq (\lambda_1 - \lambda_{s_0} + 3\ s_0\ (\varepsilon_{max} + \varepsilon_{min}))\ \|\mathbf{h}_T\|_2 \sum_{j=2,\ldots,J} \|\mathbf{h}_{T_j}\|_2. \end{aligned}$$

By Lemma [80, Lemma A.4], we get

$$\sum_{j=2,\ldots,J} \|\mathbf{h}_{T_j}\|_2 \leq \frac{\|\mathbf{h}_{T_0^c}\|_1}{\sqrt{s_0}} \tag{6.20}$$

and we can deduce that

$$\|X_T \mathbf{h}_T\|_2^2 \leq (\lambda_1 - \lambda_{s_0} + 3\ s_0\ (\varepsilon_{max} + \varepsilon_{min}))\ \|\mathbf{h}_T\|_2\ \frac{\|\mathbf{h}_{T_0^c}\|_1}{\sqrt{s_0}}. \tag{6.21}$$

Combining (6.21) with (6.19) gives

$$\|\mathbf{h}_T\|_2 \leq \frac{\lambda_1 - \lambda_{s_0} + 3\ s_0\ (\varepsilon_{max} + \varepsilon_{min})}{\lambda_{s_0} - 3\ s_0\ \varepsilon_{min}}\ \frac{\|\mathbf{h}_{T_0^c}\|_1}{\sqrt{s_0}}.$$

□

[2] The last set contains the remaining smallest terms in absolute value and may not contain s terms

6.4 Conclusion

In this paper, we established a relationship between the coherence and a weak version of the Null Space Property for design matrices in Compressed Sensing. Our approach is based on perturbation theory and no randomness assumption on the design matrix is used to establish this property. We expect that this result will be helpful to study a larger class of designs than usually done in the literature. In a future paper, we will show that such bounds can be fruitfully applied to simplify the analysis of Robust PCA.

Appendices

6.A Technical lemmæ

6.A.1 Some perturbation results

Perturbation after appending a column to a given matrix is a special type of perturbation. A survey on this topic is [50].

Background

Recall that for a matrix A in $\mathbb{R}^{n\times n}$, p_A denotes the characteristic polynomial of A.

Lemma 6.1. Cauchy's Interlacing theorem. *If $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and associated eigenvectors $\mathbf{v_1},\ldots,\mathbf{v_n}$, and $\mathbf{v} \in \mathbb{R}^n$, then*

$$p_{A+\mathbf{v}\mathbf{v}^\top}(x) = p_A(x)\left(1 - \sum_{i=1}^{n} \frac{\langle \mathbf{v}, \mathbf{u_i}\rangle^2}{x-\lambda_i}\right). \tag{6.22}$$

The previous lemma states in particular that the eigenvalues of A interlace those of $A + \mathbf{v}\mathbf{v}^\top$.

6.A.2 Appending one vector: perturbation of the smallest non zero eigenvalue

If we consider a subset T_0 of $\{1,\ldots,p\}$ and a submatrix X_{T_0} of X, the problem of studying the eigenvalue perturbations resulting from appending a column X_j to X_{T_0}, with $j \notin T_0$ can be studied using Cauchy's Interlacing Lemma as in the following result.

Lemma 6.2. *Let $T_0 \subset \{1,\ldots,p\}$ with $|T_0| = s_0$ and X_{T_0} a submatrix of X. Let $\lambda_1 \geq ... \geq \lambda_{s_0}$ be the eigenvalues of $X_{T_0}X_{T_0}^\top$. Let $\tilde{\lambda}_{s_0} \leq \lambda_{s_0}$. Assume that $\tilde{\lambda}_{s_0} < 1 - s_0\mu^2$, we have*

$$\lambda_{s_0+1}\left(X_{T_0}X_{T_0}^\top + \mathbf{X_j}\mathbf{X_j}^\top\right) \geq \tilde{\lambda}_{s_0} - \epsilon_{s_0,min}$$

with

$$\epsilon_{s_0,min} = \frac{1}{2}\left(\frac{s_0^3\ \mu^2\ \|X_{T_0}\|^2 + 4s_0^{\frac{3}{2}}\ \mu\ \|X_{T_0}\|\tilde{\lambda}_{s_0}}{2\ \left(1 - s_0\mu^2 - \tilde{\lambda}_{s_0}\right)}\right).$$

Proof. Setting $\mathbf{v} = \mathbf{X_j}$

$$A = X_{T_0} X_{T_0}^\top$$

we obtain that the smallest nonzero eigenvalue of $X_{T_0} X_{T_0}^\top + \mathbf{X_j}\mathbf{X_j}^\top$ is the smallest root $\rho_{\min}$ of

$$f(x) = 1 - \sum_{i=1}^{n} \frac{\langle \mathbf{v}, \mathbf{u_i} \rangle^2}{x - \lambda_i}.$$

Therefore, $\rho_{\min}$ is larger than the smallest positive root of

$$\tilde{f}(x) = 1 - \frac{s_0\ \gamma}{x - \tilde{\lambda}_{s_0}} - \frac{1 - s_0\mu^2}{x}$$

for any upper bound γ to $\langle \mathbf{v}, \mathbf{u_i} \rangle^2$ for $i = 1, \ldots, s_0$. Thus, we find that

$$\rho_{\min} \geq \frac{1}{2}\left(s_0(\gamma - \mu^2) + \tilde{\lambda}_{s_0} + 1 - \sqrt{s_0^2\gamma^2 + 2s_0\gamma\left(\tilde{\lambda}_{s_0} + 1 - s_0\mu^2\right) + \left(1 - s_0\mu^2 - \tilde{\lambda}_{s_0}\right)^2} \right). \tag{6.23}$$

As long as $1 - s_0\mu^2 > \lambda_{s_0}$, we have

$$\rho_{\min} \geq \frac{1}{2}\left(s_0(\gamma - \mu^2) + \tilde{\lambda}_{s_0} + 1 - \left(1 - s_0\mu^2 - \tilde{\lambda}_{s_0}\right) \sqrt{1 + \frac{s_0^2\gamma^2 + 2s_0\gamma(\tilde{\lambda}_{s_0} + 1 - s_0\mu^2)}{\left(1 - s_0\mu^2 - \tilde{\lambda}_{s_0}\right)^2}} \right).$$

Moreover, since $\sqrt{1+a} \leq 1 + \frac{1}{2}a$, we get

$$\rho_{\min} \geq \frac{1}{2}\left(s_0(\gamma - \mu^2) + \tilde{\lambda}_{s_0} + 1 - \left(1 - s_0\mu^2 - \tilde{\lambda}_{s_0}\right) \left(1 + \frac{s_0^2\gamma^2 + 2s_0\gamma(\tilde{\lambda}_{s_0} + 1 - s_0\mu^2)}{2\ \left(1 - s_0\mu^2 - \tilde{\lambda}_{s_0}\right)^2}\right) \right)$$

which gives

$$\rho_{\min} \geq \tilde{\lambda}_{s_0} - \epsilon_{s_0,min} \tag{6.24}$$

with

$$\epsilon_{s_0,min} = \frac{1}{2}\left(\frac{s_0^2\gamma^2 + 4s_0\gamma\tilde{\lambda}_{s_0}}{2\ \left(1 - s_0\mu^2 - \tilde{\lambda}_{s_0}\right)} \right).$$

Let us now find out a reasonable value of γ. Let $X_{T_0} = U_0\Sigma_0 V_0^\top$ denote the singular value decomposition of X_{T_0}. We have

$$\begin{aligned} |\langle \mathbf{X_j}, \mathbf{u_{jo}} \rangle| &= |\langle X_j, X_{T_0} V_0 \Sigma_0 \mathbf{e_{jo}} \rangle| \\ &= \|X_{T_0}^\top \mathbf{X_j}\|_2 \|V_0\Sigma_0\mathbf{e_{jo}}\|_2 \\ &\leq \sqrt{s_0}\ \mu\ \|X_{T_0}\|. \end{aligned}$$

Therefore we can take

$$\gamma = \sqrt{s_0}\ \mu\ \|X_{T_0}\|.$$

Combining this result with (6.24), we get the desired result. □

6.A.3 Appending one vector: perturbation of the largest eigenvalue

For the largest eigenvalue, we obtain

Lemma 6.3. *Let $T_0 \subset \{1, \ldots, p\}$ with $|T_0| = s_0$ and X_{T_0} a submatrix of X. Let $\lambda_1 \geq ... \geq \lambda_{s_0}$ be the eigenvalues of $X_{T_0}X_{T_0}^\top$. Let $\tilde{\lambda}_1 \geq \lambda_1$, with $\tilde{\lambda}_1 > 1$. Then, we have*

$$\lambda_1 \left(X_{T_0}X_{T_0}^\top + \mathbf{X_j}\mathbf{X_j}^\top \right) \leq \tilde{\lambda}_1 + \epsilon_{s_0,\max}.$$

with

$$\epsilon_{s_0,\max} = \frac{1}{2}\left(\frac{s_0^3\mu^2\|X_{T_0}\|^2 + 4s_0^{3/2}\mu\ \|X_{T_0}\|\tilde{\lambda}_1}{2(\lambda_1 - 1)} \right).$$

Proof. Setting $\mathbf{v} = \mathbf{X_j}$

$$A = X_{T_0}X_{T_0}^\top$$

we obtain that the largest nonzero eigenvalue of $X_{T_0}X_{T_0}^\top + \mathbf{X_j}\mathbf{X_j}^\top$ is the largest root $\rho_{\max}$ of

$$f(x) = 1 - \sum_{i=1}^{n} \frac{\langle \mathbf{v}, \mathbf{u_i} \rangle^2}{x - \lambda_i}.$$

Therefore, $\rho_{\max}$ is smaller than the largest positive root of

$$\tilde{f}(x) = 1 - \frac{s_0\ \gamma}{x - \tilde{\lambda}_1} - \frac{1}{x}$$

for any upper bound γ to $\langle \mathbf{v}, \mathbf{u_i} \rangle^2$ for $i = 1, \ldots, s_0$. Hence, we find that

$$\rho_{\max} \leq \frac{1}{2}\left(s_0\gamma + \tilde{\lambda}_1 + 1 + \sqrt{s_0^2\gamma^2 + 2s_0\gamma\left(\tilde{\lambda}_1 + 1\right) + \left(1 - \tilde{\lambda}_1\right)^2} \right). \tag{6.25}$$

Since the columns of X have unit ℓ_2-norm, we have $1 < \lambda_1$, and thus one obtains from (6.25) that

$$\rho_{\max} \leq \frac{1}{2}\left(s_0\gamma + \tilde{\lambda}_1 + 1 + \left(\tilde{\lambda}_1 - 1\right)\sqrt{1 + \frac{s_0^2\gamma^2 + 2s_0\gamma(\tilde{\lambda}_1 + 1)}{\left(\tilde{\lambda}_1 - 1\right)^2}} \right)$$

which gives

$$\rho_{\max} \leq \tilde{\lambda}_1 + \epsilon_{s_0,\max}$$

with

$$\epsilon_{s_0,\max} = \frac{1}{2}\left(\frac{s_0^2\gamma^2 + 4s_0\gamma\tilde{\lambda}_1}{2(\tilde{\lambda}_1 - 1)}\right).$$

We finally plug in the value of γ found earlier in the proof of Lemma 6.2 to get the desired result. □

6.A.4 Successive perturbations

If we append s_0 columns successively, we obtain the following result.

Lemma 6.4. *Let $T_0 \subset \{1,\ldots,p\}$ with $|T_0| = s_0$ and X_{T_0} a submatrix of X. Let $\lambda_1 \geq ... \geq \lambda_{s_0}$ be the eigenvalues of $X_{T_0}X_{T_0}^\top$. Let $\tilde{\lambda}_1 \geq \lambda_1$ and $\tilde{\lambda}_{s_0} \leq \lambda_{s_0}$. Let $T_1 \subset \{1,\ldots,p\}$ with $|T_1| = s_1$ and $T_0 \cap T_1 = \emptyset$. Assume*

1. $1 - (s_0 + s_1)\mu > \tilde{\lambda}_{s_0} > \eta$;
2. $1 < \tilde{\lambda}_1 < 2 - \eta$;
3. $s_1 < \min\left(\frac{\tilde{\lambda}_{s_0} - \eta}{\varepsilon_{min}}, \frac{2-\eta-\tilde{\lambda}_1}{\varepsilon_{max}}\right)$;

with

$$\varepsilon_{min} = \frac{1}{4}\left(\frac{s_0^3\ \mu^2\ \eta^2 + 4s_0^{3/2}\ \mu\ \eta^2}{(1 - s_0\mu^2 - \eta)}\right)$$

$$\varepsilon_{max} = \frac{1}{4}\left(\frac{(s_0+s_1)^3\mu^2(2-\eta)^2 + 4(s_0+s_1)^{3/2}\mu(2-\eta)^2}{(\tilde{\lambda}_1 - 1)}\right)$$

Then

$$\lambda_1\left(X_{T_0\cup T_1}^\top X_{T_0\cup T_1}\right) \leq \tilde{\lambda}_{s_0} - s_1\varepsilon_{min} \tag{6.26}$$

and

$$\lambda_{s_0+s_1}\left(X_{T_0\cup T_1}^\top X_{T_0\cup T_1}\right) \geq \tilde{\lambda}_1 + s_1\varepsilon_{max} \tag{6.27}$$

Proof. The proof relies on induction. First of all, note that from assumption (3)

(i) $\tilde{\lambda}_{s_0} - s_1\varepsilon_{min} > \eta$;

(ii) $\tilde{\lambda}_1 + s_1\varepsilon_{max} < 2 - \eta$.

We apply lemma 6.2 to $X_{T_0}X_{T_0}^\top + \mathbf{X_{j_1}}\mathbf{X_{j_1}}^\top$ with $j_1 \in T_1$. We have

$$\lambda_{s_0+1}(X_{T_0}X_{T_0}^\top + \mathbf{X_{j_1}}\mathbf{X_{j_1}}^\top) \geq \tilde{\lambda}_{s_0} - \varepsilon_{s_0,min}$$

with ε defined in 6.2. Since $\varepsilon_{s_0,min} \leq \varepsilon_{min}$, we get

$$\lambda_{s_0}(X_{T_0}X_{T_0}^\top) \geq \lambda_{s_0+1}(X_{T_0}X_{T_0}^\top + \mathbf{X_{j_1}}\mathbf{X_{j_1}}^\top) \geq \tilde{\lambda}_{s_0} - \varepsilon_{min}.$$

It implies by (i) that

$$1 - (s_0 + s_1)\mu > \lambda_{s_0+1}(X_{T_0}X_{T_0}^\top + \mathbf{X_{j_1}}\mathbf{X_{j_1}}^\top) > \eta.$$

Thus, the induction hypothesis is verified and we can apply Lemma 6.2 for the next step of the induction. This leads to (6.26).

For the lower bound (6.27), we have from lemma 6.A.3

$$\lambda_1(X_{T_0}X_{T_0}^\top + \mathbf{X_{j_1}}\mathbf{X_{j_1}}^\top) \leq \tilde{\lambda}_1 + \varepsilon_{s_0,max}$$

Since $\varepsilon_{s_0,max} \leq \varepsilon_{max}$, we have by (ii) that

$$1 < \lambda_1(X_{T_0}X_{T_0}^\top + \mathbf{X_{j_1}}\mathbf{X_{j_1}}^\top) < 2 - \eta.$$

We can then apply lemma 6.A.3 to the next step. The result follows by induction. □

Corollary 6.1. *Let $T_0 \subset \{1, \ldots, p\}$ with $|T_0| = s_0$ and X_{T_0} a submatrix of X. Let $\lambda_1 \geq \ldots \geq \lambda_{s_0}$ be the eigenvalues of $X_{T_0}X_{T_0}^\top$. Let $\tilde{\lambda}_1 \geq \lambda_1$ and $\tilde{\lambda}_{s_0} \leq \lambda_{s_0}$. Let $T_1 \subset \{1, \ldots, p\}$ with $|T_1| = s_1$ and $T_0 \cap T_1 = \emptyset$. Set $\eta = \frac{1}{2}$ and $s_1 = 3s_0$. Assume*

1. $1 - (s_0 + s_1)\mu > \tilde{\lambda}_{s_0} > \eta$;
2. $1 < \tilde{\lambda}_1 < 2 - \eta$;
3. $s_1 < \min\left(\frac{\tilde{\lambda}_{s_0} - \eta}{\varepsilon_{min}}, \frac{2-\eta-\tilde{\lambda}_1}{\varepsilon_{max}}\right)$;

with

$$\varepsilon_{min} = \frac{1}{4}\frac{s_0^3\ \mu^2\ /4 + s_0^{3/2}\ \mu}{\left(1 - s_0\mu^2 - \frac{1}{2}\right)}$$

$$\varepsilon_{max} = \frac{1}{4}\left(\frac{144\ s_0^4\mu^2 + 32\ s_0^{3/2}\mu(2-\eta)^2}{(\tilde{\lambda}_1 - 1)}\right)$$

Assume also

$$\mu \leq \min \left\{ \frac{1}{\sqrt{288 s_0^{5/2} \left(2 s_0^{3/2} + 1\right)}}, \frac{1}{\sqrt{\frac{3}{2} s_0^4 + 6 s_0^{5/2} + 2 s_0}} \right\}.$$

Then,

$$\lambda_1 \left(X_{T_0 \cup T_1}^\top X_{T_0 \cup T_1} \right) \leq \tilde{\lambda}_1 + 3 s_0 \ \varepsilon_{max} \tag{6.28}$$

and

$$\lambda_{s_0+s_1} \left(X_{T_0 \cup T_1}^\top X_{T_0 \cup T_1} \right) \geq \tilde{\lambda}_{s_0} - 3 s_0 \ \varepsilon_{min}. \tag{6.29}$$

Proof. Set $\eta = \frac{1}{2}$, assumption (3) writes

$$s_1 < \frac{4(\tilde{\lambda}_{s_0} - \frac{1}{2})(\frac{1}{2} - s_0 \mu^2)}{s_0^3 \mu^2 \frac{1}{4} + s_0^{3/2} \mu}$$

and

$$s_1 < \frac{4(\frac{3}{2} - \tilde{\lambda}_1)(\tilde{\lambda}_1 - 1)}{(s_0 + s_1)^3 \mu^2 \frac{9}{4} + 9 (s_0 + s_1)^{3/2} \mu}$$

which leads to

$$s_1 \left(s_0^3 \mu^2 \frac{1}{4} + s_0^{3/2} \mu \right) + 4 \left(\tilde{\lambda}_{s_0} - \frac{1}{2} \right) s_0 \mu^2 < 2 \left(\tilde{\lambda}_{s_0} - \frac{1}{2} \right)$$

and

$$s_1 \left((s_0 + s_1)^3 \mu^2 \frac{9}{4} + 9 (s_0 + s_1)^{3/2} \mu \right) < 4 \left(\frac{3}{2} - \tilde{\lambda}_1 \right) (\tilde{\lambda}_1 - 1).$$

Since $\mu < 1$

$$s_1 \left(s_0^3 \mu^2 \frac{1}{4} + s_0^{3/2} \mu^2 \right) + 4 \left(\tilde{\lambda}_{s_0} - \frac{1}{2} \right) s_0 \mu^2 < 2 \left(\tilde{\lambda}_{s_0} - \frac{1}{2} \right)$$

and

$$s_1 \left((s_0 + s_1)^3 \mu^2 \frac{9}{4} + 9 (s_0 + s_1)^{3/2} \mu^2 \right) < 4 \left(\frac{3}{2} - \tilde{\lambda}_1 \right) (\tilde{\lambda}_1 - 1)$$

The result follows by factoring out μ^2 and setting $s_1 = 3 s_0$. □

6.A.5 Bounding scalar products

Lemma 6.5. *Let $|T_0| = s_0$ and $|T_1| = s_0$, T_0, T_1 disjoint. Let $T = T_0 \cup T_1$ and T' be two disjoint subsets of $\{1, \ldots, p\}$ with $|T'| = 2s_0$. Let $\mathbf{g}$ and $\mathbf{h}$ be vectors in $\mathbb{R}^p$. Assume that*

$$\mu \leq \min \left\{ \frac{1}{\sqrt{288 s_0^{5/2} \left(2 s_0^{3/2} + 1\right)}}, \frac{1}{\sqrt{\frac{3}{2} s_0^4 + 6 s_0^{5/2} + 2 s_0}} \right\}.$$

Then,

$$|\langle X_T \mathbf{g}_T, X_{T'} \mathbf{h}_{T'} \rangle| \leq (\lambda_1 + 3\ s_0\ \varepsilon_{max})\ \|\mathbf{g}_T\|_2\ \|\mathbf{h}_{T'}\|_2. \tag{6.30}$$

Proof. Assume first that $\|\mathbf{g}_T\|_2 = \|\mathbf{h}_{T'}\|_2 = 1$. The parallelogram law now gives

$$\begin{aligned} |\langle X_T \mathbf{g}_T, X_{T'} \mathbf{h}_{T'} \rangle| &\leq \frac{1}{4} \left| \|X_T \mathbf{g}_T + X_{T'} \mathbf{h}_{T'}\|_2^2 - \|X_T \mathbf{g}_T - X_{T'} \mathbf{h}_{T'}\|_2^2 \right| \\ &\leq \frac{1}{4} \left| \|X_T \mathbf{g}_T + X_{T'} \mathbf{h}_{T'}\|_2^2 - \|X_T \mathbf{g}_T - X_{T'} \mathbf{h}_{T'}\|_2^2 \right| \end{aligned}$$

Notice that

$$\|X_T \mathbf{g}_T \pm X_{T'} \mathbf{h}_{T'}\|_2^2 = \|X_{T \cup T'} (\mathbf{g}_T \pm \mathbf{h}_{T'})\|_2^2.$$

By Corollary 6.1, we have

$$(\lambda_{s_0} - 3\ s_0\ \varepsilon_{min}) \|\mathbf{g}_T + \mathbf{h}_{T'}\|_2^2\ \leq \|\mathbf{g}_T + \mathbf{h}_{T'}\|_2^2 \|X_T \mathbf{g}_T \pm X_{T'} \mathbf{h}_{T'}\|_2^2 \leq (\lambda_1 + 3\ s_0\ \varepsilon_{max})\ \|\mathbf{g}_T + \mathbf{h}_{T'}\|_2^2.$$

From this, and the fact that g_T and h_T are unit norm, we deduce that

$$|\langle X_T \mathbf{g}_T, X_{T'} \mathbf{h}_{T'} \rangle| \leq \lambda_1 - \lambda_{s_0} + 3\ s_0\ (\varepsilon_{max} + \varepsilon_{min}).$$

The proof is completed using homogeneity. □

Chapter 7

Incoherent submatrix selection via approximate independence sets in scalar product graphs

7.1 Introduction

The goal of the present paper is to address the problem of incoherent submatrix extraction. Recall that the coherence of a matrix X with ℓ_2-normalised columns is the maximum absolute value of the scalar product of any two different columns of X. It is usually denoted by $\mu(X)$. Controlling the coherence of a matrix is of paramount importance in statistics [42], [52], [51], [177], signal processing, compressed sensing and image processing, [40], [36], [46], [145], [11], [123], [86], [2], etc. Incoherence is associated with interesting questions in combinatorics [130]. Efficient recovery of incoherent matrices, an important problem in dictionary learning, was addressed by the computer science community in [5]. Incoherence is a key assumption behind the current approaches of sparse estimation based on convex optimisation [34].

In many real life problems from statistics, and signal and image processing, the incoherence assumption breaks down [41], [2], [1], etc. Several approaches helping to get around this problem have been proposed in the literature but, to the best of our knowledge, the problem of extracting a sufficiently incoherent submatrix from a given matrix has not yet been studied in the literature. On the other hand, incoherent submatrix extraction is an important problem and a computationally efficient method for it will definitely allow to select sufficiently different features from data and make high dimensional sparse representation of the data possible in difficult settings where naive use of ℓ_1 penalised estimation was previously doomed to slow learning rates, [32], [19].

In the present paper, we propose a new approach to address the incoherent submatrix selection problem seen as a weighted version of the independent set problem in a graph. More precisely, we propose a new spectral-type estimator for the column subset selection problem and study its performance by bounding its scalar product with the indicator vector of the best column selection, in the case it is unique.

The plan of the paper is as follows. In Section 7.2, we reformulate the problem as an independent set computation problem. Section 7.3 presents a new spectral estimator of the weighted stable set and provides a theoretical error bound.

7.2 Incoherent submatrix extraction as an approximate independent set computation

The problem of extracting the largest submatrix with coherence less than a given threshold η, by appropriate column selection, can be expressed as an instance of the maximum stable set in graph theory. In order to achieve this, we associate with our matrix X a graph $\mathcal{G} = (V, E)$ as follows:

- $V = \{1, \ldots, p\}$
- E is defined by

$$(j, j') \in E \text{ if and only if } |\langle \mathbf{X}_j, \mathbf{X}_{j'} \rangle| > \eta. \tag{7.1}$$

Then, clearly, finding the largest stable set in this graph will immediately provide a submatrix with coherence less than η.

The main difficulty with the independent set approach is that it belongs to the class of NP-hard problems.

7.3 Relaxing on the sphere: a new extraction approach

In this section, we present a new spectral-type estimator for the independent set.

7.3.1 The spectral estimator

We start from the following relaxed problem

$$\max_{x \in \{0,1\}^p} \mathbf{e}^\top \mathbf{x} \quad \text{s.t.} \quad \mathbf{x}^\top M \mathbf{x} \leq r. \tag{7.2}$$

The approach of the previous section was addressing the special case corresponding to the value $\lambda = 0$ and M chosen as e.g. the adjacency matrix of $\mathcal{G}$. As in the previous section, we can reformulate this problem using a binary ± 1 variable $\mathbf{z}$ as follows:

$$\max_{\mathbf{z} \in \{-1,1\}^p} \quad \mathbf{e}^\top \frac{\mathbf{z} + \mathbf{e}}{2} \quad \text{s.t.} \quad \frac{1}{4} (\mathbf{z} + \mathbf{e})^\top M (\mathbf{z} + \mathbf{e}) \leq r. \tag{7.3}$$

The spectral estimator is the vector obtained as the solution of the following problem.

$$\max_{\|\mathbf{z}\|_2^2 = p} \quad \mathbf{e}^\top \frac{\mathbf{z} + \mathbf{e}}{2} \quad \text{s.t.} \quad \frac{1}{4} (\mathbf{z} + \mathbf{e})^\top M (\mathbf{z} + \mathbf{e}) \leq r. \tag{7.4}$$

It is equivalent to the penalised version

$$\min_{\|\mathbf{z}\|_2^2=p} \quad -\mathbf{e}^\top \frac{\mathbf{z}+\mathbf{e}}{2} + \lambda \left(\frac{1}{4} (\mathbf{z}+\mathbf{e})^\top M(\mathbf{z}+\mathbf{e}) - r \right) \tag{7.5}$$

for some Lagrange multiplier $\lambda > 0$ [1].

7.3.2 Theoretical guarantees

In this section, we provide our main theoretical result concerning the performance of our approach.

Theorem 7.1. *Let $\boldsymbol{\rho}^*$ denote the indicator vector of the maximal independent set defined by*

$$\max_{\boldsymbol{\rho}\in\{0,1\}^p} \ \mathbf{e}^\top \boldsymbol{\rho} \quad s.t. \quad \boldsymbol{\rho}^\top M \boldsymbol{\rho} = 0.$$

Let x_2^* be solution of (7.5). Let $\delta > 0$ be such that $M_\delta = M + \delta I$ is positive definite. Let λ_1 be the smallest eigenvalue of M_δ and $\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_p$ be the pairwise orthogonal, unit-norm eigenvectors of M_δ. Set

$$\mathbf{q}_1 = \frac{1}{\sqrt{p}} M_\delta \mathbf{e} \quad \text{and} \quad \mathbf{q}_2 = -\frac{1}{\sqrt{p}} \Big(\frac{(1+\delta)}{\lambda} \ \mathbf{e} - M_\delta \mathbf{e} \Big)$$

and set

$$\gamma_{k,i} = \boldsymbol{\phi}_i^\top \mathbf{q}_k$$

for $k = 1, 2$ and $i = 1, \ldots, p$. Then,

$$\|\boldsymbol{\rho}^* - \mathbf{x_2}^*\|_\infty \leq \sqrt{p} \left(\frac{(1+\delta)}{\lambda(\lambda_1 - \mu_2)} + \frac{\|\gamma_1\|_2 \ r^*}{(\lambda_1 - \mu_1)(\lambda_1 - \mu_2)} \right)^2,$$

with r^* given by

$$r^* = (\lambda_p - \mu_1)\phi \left(p \ \frac{\gamma_{1,\max}^2}{\gamma_{1,\min}^2} \frac{\frac{(1+\delta)^2}{\lambda^2} + \frac{2}{\sqrt{p}} \frac{1+\delta}{\lambda} \mathbf{e}^\top M_\delta \mathbf{e}}{2 \ \|\gamma_2\|_2^2} \right)$$

where ϕ denotes the inverse function of $x \mapsto x/(1+x)^3$.

Proof. The proof will consist of three steps. The first step re-expresses the problem as the one of minimising the distance to the oracle plus a linear penalisation term. The second step identifies the oracle as the solution to a perturbed problem. The third step then uses a perturbation result proved in the Appendix.

[1] here, positivity is trivial

first step. Let us now note that since $\|\mathbf{z}\|_2^2 = p$, we may incorporate any multiple of the term $\mathbf{z}^\top\mathbf{z} - p$ into the objective function and obtain

$$\min_{\|\mathbf{z}\|_2^2=p} \quad -\mathbf{e}^\top \frac{\mathbf{z}+\mathbf{e}}{2} + \lambda \left(\frac{1}{4}\,(\mathbf{z}+\mathbf{e})^\top M(\mathbf{z}+\mathbf{e}) - r\right) + \frac{1}{4}\,\delta(\mathbf{z}^\top\mathbf{z} - p),$$

without changing the solution. Using the fact that

$$\begin{aligned}\mathbf{z}^\top\mathbf{z} - p &= (\mathbf{z}+\mathbf{e}-\mathbf{e})^\top(\mathbf{z}+\mathbf{e}-\mathbf{e}) - p\\ &= (\mathbf{z}+\mathbf{e})^\top(\mathbf{z}+\mathbf{e}) - 2\mathbf{e}^\top(\mathbf{z}+\mathbf{e})\end{aligned}$$

we obtain the equivalent problem

$$\min_{\|\mathbf{z}\|_2^2=p} \quad -(1+\delta)\;\mathbf{e}^\top\frac{\mathbf{z}+\mathbf{e}}{2} + \frac{\lambda}{4}\,(\mathbf{z}+\mathbf{e})^\top(M+\delta I)(\mathbf{z}+\mathbf{e}). \tag{7.6}$$

Set $M_\delta = M + \delta$. We can now expand the term

$$\begin{aligned}\frac{1}{4}\,(\mathbf{z}+\mathbf{e})^\top M_\delta(\mathbf{z}+\mathbf{e}) &= \left(\frac{\mathbf{z}+\mathbf{e}}{2} - \boldsymbol{\rho}^* + \boldsymbol{\rho}^*\right)^\top M_\delta \left(\frac{\mathbf{z}+\mathbf{e}}{2} - \boldsymbol{\rho}^* + \boldsymbol{\rho}^*\right)\\ &= \left(\frac{\mathbf{z}+\mathbf{e}}{2} - \boldsymbol{\rho}^*\right)^\top M_\delta \left(\frac{\mathbf{z}+\mathbf{e}}{2} - \boldsymbol{\rho}^*\right)\\ &\quad + 2\left(\frac{\mathbf{z}+\mathbf{e}}{2}\right)^\top M_\delta\boldsymbol{\rho}^*,\end{aligned}$$

where we used the fact that ${\boldsymbol{\rho}^*}^\top M\boldsymbol{\rho}^* = 0$ in the last equality. Therefore, (7.6) is equivalent to

$$\min_{\|\mathbf{z}\|_2^2=p} \quad \left(2M_\delta\boldsymbol{\rho}^* - \frac{(1+\delta)}{\lambda}\;\mathbf{e}\right)^\top\left(\frac{\mathbf{z}+\mathbf{e}}{2}\right) + \left(\frac{\mathbf{z}+\mathbf{e}}{2} - \boldsymbol{\rho}^*\right)^\top M_\delta\left(\frac{\mathbf{z}+\mathbf{e}}{2} - \boldsymbol{\rho}^*\right). \tag{7.7}$$

Second step. When $1/\lambda = 0$, the solution to (7.7) is readily seen to be equal to the oracle $\boldsymbol{\rho}^*$.

Third step. We will now use a perturbation result proved in Lemma 7.3. For this purpose, we first make a change of variable in order to transform the problem into an optimisation problem on the unit sphere. Let $\tilde{\mathbf{z}} = 1/\sqrt{p}\mathbf{z}$. Then problem (7.7) is equivalent to

$$\begin{aligned}\min_{\|\tilde{\mathbf{z}}\|_2^2=1} \quad &\left(2M_\delta\boldsymbol{\rho}^* - \frac{(1+\delta)}{\lambda}\;\mathbf{e}\right)^\top\left(\frac{\sqrt{p}\tilde{\mathbf{z}}+\mathbf{e}}{2}\right)\\ &+ \left(\frac{\sqrt{p}\tilde{\mathbf{z}}+\mathbf{e}}{2} - \boldsymbol{\rho}^*\right)^\top M_\delta\left(\frac{\sqrt{p}\tilde{\mathbf{z}}+\mathbf{e}}{2} - \boldsymbol{\rho}^*\right).\end{aligned}$$

This is equivalent to solving the problem

$$\min_{\|\tilde{\mathbf{z}}\|_2=1} \quad \frac{1}{2}\tilde{\mathbf{z}}^\top Q\tilde{\mathbf{z}} - q^\top\tilde{\mathbf{z}}. \tag{7.8}$$

with

$$Q = M_\delta \quad \text{and} \quad \mathbf{q} = -\frac{1}{\sqrt{p}}\Big(\frac{(1+\delta)}{\lambda}\,\mathbf{e} - M_\delta \mathbf{e}\Big).$$

Let $\lambda_1 \leq \ldots \leq \lambda_p$ be the eigenvalues of Q and $\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_p$ be associated pairwise orthogonal, unit-norm eigenvectors. Set

$$\mathbf{q}_1 = \frac{1}{\sqrt{p}} M_\delta \mathbf{e} \quad \text{and} \quad \mathbf{q}_2 = -\frac{1}{\sqrt{p}}\Big(\frac{(1+\delta)}{\lambda}\,\mathbf{e} - M_\delta \mathbf{e}\Big)$$

and set

$$\gamma_{k,i} = \boldsymbol{\phi}_i^\top \mathbf{q}_k$$

for $k = 1, 2$ and $i = 1, \ldots, p$. Thus, we have

$$\|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\|_2 = \frac{(1+\delta)}{\lambda}.$$

and

$$\|\boldsymbol{\gamma}_1^2 - \boldsymbol{\gamma}_2^2\|_1 \leq \frac{(1+\delta)^2}{\lambda^2} + \frac{2}{\sqrt{p}}\frac{1+\delta}{\lambda}\mathbf{e}^\top M_\delta \mathbf{e}.$$

Combining these bounds with Lemma 7.3, we obtain the announced result. □

7.4 Conclusion and future works

In the present paper, we proposed an alternative approach to the problem of column selection, viewed as quadratic binary maximisation problem. We studied the approximation error of the solution to the easier spherically constrained quadratic problem obtained by relaxing the binary constraints.

Future work will consist in further exploring the quality of the bound obtained in Theorem 7.1. In particular, we will try to clarify for which types of graphs the error bound can be small, i.e. $|\lambda_1 - \mu_1|$ is bounded from below by an appropriate function of p. Our next plans will also address practical assessment of the efficiency of the method.

Appendices

7.A Minimising quadratic functionals on the sphere

7.A.1 A semi-explicit solution

The following result can be found in [99].

Lemma 7.1. *For $Q \in \mathbb{S}_p$ and $\mathbf{q} \in \mathbb{R}^p$, consider the following quadratic programming problem over the sphere:*

$$\min_{\|\mathbf{x}\|_2=1} \quad \frac{1}{2}\mathbf{x}^\top Q\mathbf{x} - \mathbf{q}^\top \mathbf{x}. \tag{7.9}$$

Let $\lambda_1 \leq \ldots \leq \lambda_p$ be the eigenvalues of Q and $\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_p$ be associated pairwise orthogonal, unit-norm eigenvectors. Let $\gamma_{k,i} = \mathbf{q}^\top \boldsymbol{\phi}_i$, $i = 1,\ldots,p$. Let $\mathcal{E}_1 = \{i \text{ s.t. } \lambda_i = \lambda_1\}$ and $\mathcal{E}_+ = \{i \text{ s.t. } \lambda_i > \lambda_1\}$. Then, x^* is a solution if and only if

$$x^* = \sum_{i=1}^{p} c_i^* \boldsymbol{\phi}_i$$

and

1. degenerate case: *If $\gamma_i = 0$ for all $i \in \mathcal{E}_1$ and*

$$\sum_{i\in\mathcal{E}_+} \frac{\gamma_i^2}{(\lambda_i - \lambda_1)^2} \leq 1.$$

   then $c_i^* = \gamma_i/(\lambda_i - \lambda_1)$, $i \in \mathcal{E}_1$ and c_i^*, $i \in \mathcal{E}_1$ are arbitrary under the constraint that $\sum_{i\in\mathcal{E}_1} \quad c_i^{*^2} = 1 - \sum_{i\in\mathcal{E}_+} \quad c_i^{*^2}$.

2. non-degenerate case: *If not in the degenerate case, $c_i^* = \gamma_i/(\lambda_i - \mu)$, $i = 1,\ldots,n$ for $\mu > -\lambda_1$ which is a solution of*

$$\sum_{i=1,\ldots,n} \frac{\gamma_i^2}{(\lambda_i - \mu)^2} = 1. \tag{7.10}$$

Moreover, we have the following useful result.

Corollary 7.1. *If Q is positive definite, and $\sum_{i=1,\ldots,p} \gamma_i^2/\lambda_i^2 < 1$, then $0 < \mu < \lambda_1$.*

Proof. This follows immediately from the intermediate value theorem. □

7.A.2 Bounds on μ

From (7.10), we can get the following easy bounds on μ.

Lemma 7.2. *Let* $\gamma_{\min} = \min_{i=1}^{p} \gamma_i$ *and* $\gamma_{\max} = \max_{i=1}^{p} \gamma_i$. *Then, we have*

$$p\gamma_{\max}^2 \geq \max_{i=1}^{p} \{(\lambda_i - \mu)^2\} \geq p\gamma_{\min}^2. \tag{7.11}$$

and

$$\gamma_{\min}^2 \leq \min_{i=1}^{p} \{(\lambda_i - \mu)^2\} \leq \|\gamma\|_2^2 \tag{7.12}$$

Proof. The proof is divided into three parts, corresponding to each (double) inequality.

Proof of (7.11): We have

$$\begin{aligned}\max_{i=1}^{p} \frac{\gamma_{\max}^2}{(\lambda_i - \mu)^2} &\geq \max_{i=1}^{p} \frac{\gamma_i^2}{(\lambda_i - \mu)^2} \\ &\geq \frac{1}{p} \sum_{i=1}^{p} \frac{\gamma_i^2}{(\lambda_i - \mu)^2} \\ &= \frac{1}{p}.\end{aligned}$$

This immediately gives $p\gamma_{\max} \geq \max_{i=1}^{p} \{(\lambda_i - \mu)^2\}$. On the one hand, we have

$$1 = p \sum_{i=1,\ldots,p} \frac{\gamma_i^2}{(\lambda_i - \mu)^2} \geq \frac{p\gamma_{\min}^2}{\max_{i=1}^{p}\{(\lambda_i - \mu)^2\}}.$$

Therefore, we get $\max_{i=1}^{p}\{(\lambda_i - \mu)^2\} \geq p\ \gamma_{\min}^2$. On the other hand, we have

Proof of (7.12):

$$\frac{\gamma_i^2}{(\lambda_i - \mu)^2} \leq 1$$

which gives

$$(\lambda_i - \mu)^2 \geq \gamma_i^2$$

for $i = 1, \ldots, p$. Thus, the lower bound follows. For the other bound, since

$$\sum_{i=1}^{p} \frac{\gamma_i^2}{(\lambda_i - \mu)^2} = 1, \tag{7.13}$$

we get

$$1 \leq \sum_{i=1}^{p} \frac{\gamma_i^2}{(\lambda_i - \mu)^2} \leq \frac{\|\boldsymbol{\gamma}\|_2^2}{\min_{i=1}^{p}\ (\lambda_i - \mu)^2}$$

and the proof in completed. □

7.A.3 ℓ_∞ perturbation of the linear term

We now consider the problem of controlling the solution under perturbation of $\mathbf{q}$.

Lemma 7.3. *Consider the two quadratic programming problems over the sphere:*

$$\min_{\|\mathbf{x}\|_2=1} \quad \frac{1}{2}\mathbf{x}^\top Q\mathbf{x} - \mathbf{q}_k^\top \mathbf{x}, \tag{7.14}$$

for $k = 1, 2$. Assume that the solution to (7.14) is non-degenerate in both cases $k = 1, 2$ and let $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ be the corresponding solutions. Assume further that $\sum_{i=1,\dots,n} \gamma_{k,i}^2/\lambda_i^2 < 1$, $k = 1, 2$. Let ϕ denote the inverse function of $x \mapsto x/(1+x)^3$. Then, we have

$$\|\mathbf{x}_1^* - \mathbf{x}_2^*\|_\infty \leq \sqrt{p}\left(\frac{\|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\|_2}{(\lambda_1 - \mu_2)} + \frac{\|\boldsymbol{\gamma}_1\|_2\ r^*}{(\lambda_1 - \mu_1)(\lambda_1 - \mu_2)}\right)^2,$$

with r^* given by

$$r^* = (\lambda_p - \mu_1)\phi\left(p\ \frac{\gamma_{1,\max}^2}{\gamma_{1,\min}^2}\frac{\|\boldsymbol{\gamma}_1^2 - \boldsymbol{\gamma}_2^2\|_1}{2\ \|\boldsymbol{\gamma}_2\|_2^2}\right)$$

Proof. Let Φ denote the matrix whose columns are the eigenvectors of A. More precisely, $\lambda_1 \leq \cdots \leq \lambda_p$ and let ϕ_i be an eigenvector associated with λ_i, $i = 1, \dots, p$. Let $\gamma_i = \mathbf{q}^\top \boldsymbol{\phi}_i$, $i = 1, \dots, p$. Let c_1^* (resp. c_2^*) be the vector of coefficients of x_1^* (resp. x_2^*) in the eigenbasis of A. For each $k = 1, 2$, there exists a real μ_k such that

$$c_{k,i}^* = \frac{\gamma_{k,i}}{(\lambda_i - \mu_k)},$$

$i = 1, \dots, p$ for $\mu_k > -\lambda_1$ which is a solution of

$$\sum_{i=1}^p \frac{\gamma_{k,i}^2}{(\lambda_i - \mu)^2} = 1.$$

Now, apply Neuberger's Theorem 7.2 to obtain an estimation of $|\mu_1 - \mu_2|$ as a function of $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$. For this purpose, set

$$F(\mu) = \sum_{i=1}^p \frac{\gamma_{2,i}^2}{(\lambda_i - \mu)^2} - 1, \ i.e. \quad F'(\mu) = 2\sum_{i=1}^p \frac{\gamma_{2,i}^2}{(\lambda_i - \mu)^3}.$$

Now, we need to find the smallest value of r such that, for all $\mu \in B(\mu_1, r)$, we need to find a number $h \in \bar{B}(0, r)$ such that

$$h = F'(\mu)^{-1}\ F(\mu_1)$$

We therefore have that

$$h = \frac{\sum_{i=1}^p \frac{\gamma_{2,i}^2}{(\lambda_i - \mu_1)^2} - 1}{2\ \sum_{i=1}^p \frac{\gamma_{2,i}^2}{(\lambda_i - \mu)^3}} = \frac{\sum_{i=1}^p \frac{\gamma_{1,i}^2}{(\lambda_i - \mu_1)^2} - 1 + \sum_{i=1}^p \frac{\gamma_{2,i}^2 - \gamma_{1,i}^2}{(\lambda_i - \mu_1)^2}}{2\ \sum_{i=1}^p \frac{\gamma_{2,i}^2}{(\lambda_i - \mu)^3}}$$

and since

$$\sum_{i=1}^{p} \frac{\gamma_{1,i}^2}{(\lambda_i - \mu_1)^2} = 1,$$

we have

$$h \leq \frac{(\min_{i=1}^{p} \{(\lambda_i - \mu_1)^2\})^{-1} \ \|\boldsymbol{\gamma}_1^2 - \boldsymbol{\gamma}_2^2\|_1}{2 \ \|\boldsymbol{\gamma}_2\|_2^2 \ (\max\{(\lambda_i - \mu)^3\})^{-1}}$$

where $\cdot^2$ is to be understood component-wise. Moreover, since $\sum_{i=1,\ldots,p} \gamma_{k,i}^2/\lambda_i^2 < 1$, $k = 1, 2$,

$$\max\{(\lambda_i - \mu)^3\} = (\lambda_p - \mu_1 + r)^3 \text{ and } \min_{i=1}^{p}\{(\lambda_i - \mu_1)^2\} = (\lambda_1 - \mu_1)^2.$$

Thus, for $r > 0$ such that

$$r \geq \frac{\|\boldsymbol{\gamma}_1^2 - \boldsymbol{\gamma}_2^2\|_1 \ (\lambda_p - \mu_1 + r)^3}{2 \ \|\boldsymbol{\gamma}_2\|_2^2 \ (\lambda_1 - \mu_1)^2},$$

we get from Theorem 7.2 that there exists a solution to the equation $F(u) = 0$ inside the ball $\bar{B}(\mu_1, r)$. Make the change of variable

$$r = \alpha(\lambda_p - \mu_1)$$

and obtain that we need to find $\alpha \in (0, 1)$ such that

$$\frac{\alpha}{(1+\alpha)^3} \geq \frac{\|\boldsymbol{\gamma}_1^2 - \boldsymbol{\gamma}_2^2\|_1 \ (\lambda_n - \mu_1)^2}{2 \ \|\boldsymbol{\gamma}_2\|_2^2 \ (\lambda_1 - \mu_1)^2}.$$

Lemma 7.2 now gives

$$\frac{(\lambda_n - \mu_1)^2}{(\lambda_1 - \mu_1)^2} \leq p \ \frac{\gamma_{1,\max}^2}{\gamma_{1,\min}^2}$$

from which we get that the value r^* of r given by

$$r^* = (\lambda_p - \mu_1)\phi\left(p \ \frac{\gamma_{1,\max}^3}{\gamma_{1,\min}^2} \frac{\|\boldsymbol{\gamma}_1^2 - \boldsymbol{\gamma}_2^2\|_1}{2 \ \|\boldsymbol{\gamma}_2\|_2^2}\right)$$

is admissible, for $\|\boldsymbol{\gamma}_1^2 - \boldsymbol{\gamma}_2^2\|_1$ such that the term involving ϕ is less than one.

$$\begin{aligned}\frac{\gamma_{1,i}}{(\lambda_i - \mu_1)} - \frac{\gamma_{2,i}}{(\lambda_i - \mu_2)} &= \frac{\gamma_{1,i}(\lambda_i - \mu_1 + \mu_1 - \mu_2) - \gamma_{2,i}(\lambda_i - \mu_1)}{(\lambda_i - \mu_1)(\lambda_i - \mu_2)} \\ &= \frac{(\gamma_{1,i} - \gamma_{2,i})}{\lambda_i - \mu_2} + \frac{\gamma_{1,i}(\mu_1 - \mu_2)}{(\lambda_i - \mu_1)(\lambda_i - \mu_2)}.\end{aligned}$$

Therefore,

$$\|c_1^* - c_2^*\|_2^2 \leq \left(\frac{\|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\|_2}{(\lambda_1 - \mu_2)} + \frac{\|\boldsymbol{\gamma}_1\|_2 \; |\mu_1 - \mu_2|}{(\lambda_1 - \mu_1)(\lambda_1 - \mu_2)} \right)^2 .$$

Finally, using that $|\mu_1 - \mu_2| \leq r^*$, we get

$$\|c_1^* - c_2^*\|_2 \leq \left(\frac{\|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\|_2}{(\lambda_1 - \mu_2)} + \frac{\|\boldsymbol{\gamma}_1\|_2 \; r^*}{(\lambda_1 - \mu_1)(\lambda_1 - \mu_2)} \right)^2 ,$$

which gives

$$\|x_1^* - x_2^*\|_\infty \leq \sqrt{p} \left(\frac{\|\gamma_1 - \gamma_2\|_2}{(\lambda_1 - \mu_2)} + \frac{\|\gamma_1\|_2 \; r^*}{(\lambda_1 - \mu_1)(\lambda_1 - \mu_2)} \right)^2 ,$$

as announced. □

7.A.4 Neuberger's theorem

In this subsection, we recall Neuberger's theorem.

Theorem 7.2. *Suppose that $r > 0$, that $\mathbf{x} \in \mathbb{R}^p$, and that F is a continuous function from $\bar{B}(\mathbf{x}, r)$ to $\mathbb{R}^m$ with the property that for each y in $B(\mathbf{x}, r)$, there is an $\mathbf{h}$ in $\bar{B}(\mathbf{0}, r)$ such that*

$$\lim_{t \to 0+} \frac{(F(\mathbf{y} + t\mathbf{h}) - F(\mathbf{y}))}{t} = -F(\mathbf{x}). \tag{7.15}$$

Then, there exists $\mathbf{u}$ in $\bar{B}(\mathbf{x}, r)$ such that $F(\mathbf{u}) = \mathbf{0}$.

Chapter 8

Average performance analysis of the projected gradient method for online PCA

8.1 Introduction

8.1.1 Background

Principal Component Analysis (PCA) is a paramount tool in an amazingly wide scope of applications. PCA belongs to the small list of algorithms which are extensively used in data science, medicine, finance, machine learning, etc. and the list is almost infinite. PCA is one of the basic blocks in the Geometric Science of Information. Computing singular/eigenvectors easily provides nonlinear embedding of data living on low dimensional manifold in a straightforward manner [9]. The other main geometric aspect of PCA lies in the fact that eigenvectors belong to the sphere and orthogonal families of eigenvectors belong to the Stiefel manifold, an information that we should take into account when computing these objects.

In the era of Big Data though, computing a set of singular vectors might turn to be a formidable task to achieve in practice since in many cases, one is not even able to store the data matrix itself in the RAM, not even mentioning running an algorithm on it. In the recent years, the need to handle massive datasets has revived a tremendous soaring of online techniques and algorithms which incorporate the data in an incremental fashion. Online convex optimisation is now a thriving field for dozens of important contributions a year, and a remarkable impact on the way statistical estimation and machine learning is undertaken in practice [101, 161, 150]. On the other hand, however, PCA lives in yet another realm, which cannot be directly reached using the techniques recently developed for convex optimisation. PCA can be performed using optimisation over the sphere and online versions of this nonconvex optimisation problem. Online or stochastic version of PCA have been extensively studied quite recently; see in particular the review [44] for a thorough analysis of the practical performance of online methods for PCA. On the theoretical side, [151, 152, 113, 4] propose very interesting results about the behaviour of stochastic gradient

type algorithms with different implementation details and under various assumptions. In particular, [152] provides a very elegant approach to the analysis of the stochastic projected gradient descent without any assumption on the spectral gap between the largest eigenvalue and the second largest eigenvalue.

8.1.2 Our contribution

The goal of the present paper is to study the online version of the stochastic gradient algorithm for PCA. In the setting we are interested in, the entries of the matrix we want to employ PCA on are observed online, i.e. one empirical correlation coefficient at a time. Our two main contributions are the following.

- We extend the analysis presented in [152] to the online setting. In particular, we obtain a precise control on the average performance of the online method which does not depend on the separation between the first and the second eigenvalue.
- We provide a practical method to tune the learning rate, i.e. the step-size of the gradient algorithm, based on a recent version proposed in [121] of the Hedge Algorithm [88].

8.1.3 Organisation of the paper

Our main results are presented in Section 8.2 where the algorithm is described and our main theorem is given. The proof of our main theorem is exposed in Section 8.3. Implementation and numerical experiments are given in Section 8.4. In particular, a simple method for choosing the learning rate is described in Section 8.4.1. The technical lemmæ which are used in the proof of Section 8.3 are gathered in Section 8.A at the end of the paper.

8.2 Main results

8.2.1 Presentation of the problem and prior result

We use bold-faced letters to denote vectors, and capital letters to denote matrices unless specified otherwise. Given a matrix A, we denote by $A^\top$ its transpose matrix, $\|A\|$ its spectral norm and $\|A\|_{1\to 2} = \max \|\mathbf{A}_j\|_2$ the maximum ℓ_2 norm of its column. For a vector $\mathbf{v}$, we denote by $\mathbf{v}^\top$ its transpose. Moreover $(\mathbf{e}_i)_i$ denote the canonical basis of $\mathbb{R}^d$. The optimisation problem can be written

$$\min_{\mathbf{w}:\|\mathbf{w}\|=1} -\mathbf{w}^\top A\mathbf{w}, \tag{8.1}$$

where $d > 1$ and A is a symmetric positive semi-definite matrix supposed unknown. We suppose that we have access to a stream of i.i.d. matrices A_t defined as

$$A_t = d^2\ A_{i_t,j_t}\ \mathbf{e}_{i_t}\mathbf{e}_{j_t}^\top \tag{8.2}$$

and (i_t, j_t) is drawn uniformly at random from $\{1, \ldots, n\}^2$. It is easily seen that $\mathbb{E}[A_t] = A$, therefore each matrix A_t can be seen as a properly rescaled noiseless random component of A. It can be readily seen that any leading eigenvector of A is a solution of the optimisation problem.

8.2.2 The stochastic projected gradient algorithm

Given a symmetric matrix $A \in \mathbb{R}^{d\times d}$, the projected gradient algorithm writes

$$\mathbf{w}_{t+1} = (I + \eta A)\mathbf{w}_t / \|(I + \eta A)\mathbf{w}_t\|_2 \tag{8.3}$$

where η is a step-size parameter and $\mathbf{w}_0$ is the initial estimate for a leading eigenvector of A. This algorithm correspond to initialising at $\mathbf{w}_0$ then make a gradient step at each iteration followed by a projection into the unit sphere. However, since A is unknown, the stochastic gradient we will study in this paper is simply defined as

$$\mathbf{w}_{t+1} = (I + \eta A_t)\mathbf{w}_t / \|(I + \eta A_t)\mathbf{w}_t\|_2 \tag{8.4}$$

obtained by replacing A with the random matrix A_t. Since the projection on the unit sphere is a rescaling operation which is commutative with respect to the matrix product, we can leave the projection operation to the end. That is, for our analysis, it is enough to consider the equivalent algorithm which only performs projection at the end:

- Initialise $\mathbf{w}_0$ on a unit sphere,
- Perform $T > 0$ stochastic gradient step : $\mathbf{w}_{t+1} = (I + \eta A_t)\mathbf{w}_t$
- Return $\mathbf{w}_T / \|\mathbf{w}_T\|_2$.

Since our work is based on the analysis of Shamir [152], it is only fair we recall its setting and main result to highlight the differences between both approach. In [152], the stream of i.i.d. matrices A_t are also supposed positive semidefinite. In which case, the following theorem holds

Theorem 8.1. *Suppose that for some leading eigenvector* $\mathbf{v}$ *of* A, $\frac{1}{p} < \langle \mathbf{w}_0, \mathbf{v}\rangle^2$ *for some* $p > 0$ *and that for some* $b \geq 1$, *both* $\|A_t\|/\|A\|$ *and* $\|A_t - A\|/\|A\|$ *are at most* b *with probability 1. Then, if we run the algorithm* (8.4) *for* T *itération with* $\eta = \frac{1}{b\sqrt{pT}}$, *then with probability at least* $\frac{1}{cp}$, *the return* $\mathbf{w}_T$ *satisfies*

$$1 - \frac{\mathbf{w}_T^\top A \mathbf{w}_T}{\|A\|} \leq c' \frac{\log(T) b\sqrt{p}}{\sqrt{T}}, \tag{8.5}$$

where c *and* c' *are positive constants.*

Note that our setting does not fit the hypothesis for which this theorem holds. In fact, the matrices A_t in the online setting are not positive semidefinite, otherwise the matrix A is necessarily a non-negative matrix.

8.2.3 Main theorem

Without loss of generality, we will assume that $\|A\| = 1$. We wish to show that for $\varepsilon > 0$, the returned $\mathbf{w}_T$ satisfies

$$1 - \mathbf{w}_T A \mathbf{w}_T \leq \varepsilon \tag{8.6}$$

in expectation when η and T satisfies some explicit conditions. Since $\|\mathbf{w}_T\|_2 = 1$, this is equivalent to showing that $\mathbf{w}_T^\top((1-\varepsilon)I - A)\mathbf{w}_T \leq 0$.

Theorem 8.2. *Let $\varepsilon > 0$ and assume that $0 < \frac{1}{p} < \langle \mathbf{w}_0, \mathbf{v}\rangle^2$ for a leading eigenvector $\mathbf{v}$ of A. Define*

$$V_T = \mathbf{w}_0^\top \prod_{i=T}^{1} (I + \eta A_i)^\top ((1-\epsilon)I - A) \prod_{i=1}^{T} (I + \eta A_i)\mathbf{w}_0. \tag{8.7}$$

Then for T satisfying

$$T > \max\left(\frac{4p^2d^2}{\varepsilon}, \frac{\log 4p\varepsilon^{-1}}{\log\left(1 + \frac{\varepsilon}{pd^2}\right)}\right), \tag{8.8}$$

and $\eta = \frac{\varepsilon}{4pd^2}$, it holds that

$$\mathbb{E}[V_T] \leq -\frac{\varepsilon}{4p}(1 + 2\eta)^T. \tag{8.9}$$

Since $V_T = \|\mathbf{w}_T\|_2^2 \mathbf{w}_T^\top((1-\varepsilon)I - A)\mathbf{w}_T$, the theorem implies the desired result.

8.3 Proof of the Theorem 8.9

In this section, we prove our main result, namely Theorem 8.2. Define

$$\mathbf{B}_T = \prod_{i=T}^{1} (I + \eta A_i)^\top ((1-\epsilon)I - A) \prod_{i=1}^{T} (I + \eta A_i) \tag{8.10}$$

so that $V_T = \mathbf{w}_0^\top B_T \mathbf{w}_0$.

Lemma 8.1. *We have that*

$$\begin{aligned} \mathbb{E}[B_T] = \mathbb{E}[B_{T-1}] + \eta\left(A^\top \mathbb{E}[B_{T-1}] + \mathbb{E}[B_{T-1}]A\right) \\ + \eta^2 d^2 \mathrm{diag}\left(A^\top \mathrm{diag}(\mathbb{E}[B_{T-1}])A\right). \end{aligned} \tag{8.11}$$

Proof. Expand the recurrence relationship and take the expectation. Finally use Lemma 8.2 to obtain the last term of the inequality. □

Expanding the recurrence in Lemma 8.1, we have

$$\mathbb{E}[V_T] \leq \mathbf{w}_0^\top (I + 2\eta A)^\top ((1-\varepsilon)I - A)\mathbf{w}_0 + \eta^2 d^2 \sum_{i=1}^{T} (1+2\eta)^{T-i} \|\mathrm{diag}(\mathbb{E}[B_{i-1}])\| \|\mathbf{w}_0\|_2^2. \tag{8.12}$$

where the last term was obtained by using inequality (8.28) and $\|A\|_{1\to 2} \leq 1$. Using an eigendecomposition of A and $\|\mathbf{w}_0\|_2^2 = 1$ gives

$$\mathbb{E}[V_T] \leq \sum_{j=1}^{d} (1+2\eta s_j)^T (1-\varepsilon-s_j) w_{0,j}^2 + \eta^2 d^2 \sum_{i=1}^{T} (1+2\eta)^{T-i} \|\mathrm{diag}(\mathbb{E}[B_{i-1}])\|. \tag{8.13}$$

where $s_1 > \cdots > s_d$ denote the eigenvalues of A and $w_{0,j} = \langle \mathbf{w}_0, \mathbf{v}_j \rangle$ denotes the $j-th$ component of $\mathbf{w}_0$ in the basis of the eigenvectors of A. Since $s_1 = 1$, this inequality rewrites

$$\mathbb{E}[V_T] \leq -\varepsilon(1+2\eta)^T w_{0,1}^2 + \sum_{j=2}^{d} (1+2\eta s_j)^T (1-\varepsilon-s_j) w_{0,j}^2 + \eta^2 d^2 \sum_{i=1}^{T} (1+2\eta)^{T-i} \|\mathrm{diag}(\mathbb{E}[B_{i-1}])\|. \tag{8.14}$$

Now, we've identified a negative term $-\varepsilon(1+2\eta)^T w_{0,1}^2$ that we want to dominate the positive terms with. The $w_{0,j}^2$ sums to $1 - w_{0,1}^2$. Therefore the sum $\sum_{j=2}^{d} (1+2\eta s_j)^T (1-\varepsilon-s_j) w_{0,j}^2$ is less than $\max_{s\in[0,]} (1+2\eta s)^T (1-\varepsilon-s)$. Lemma 8.7 gives a bound on this maximum. In consequence, we have the following inequality

$$\mathbb{E}[V_T] \leq -\varepsilon(1+2\eta)^T w_{0,1}^2 + (1 + \frac{(1+2\eta(1-\varepsilon))^T}{\eta(T+1)}) + \eta^2 d^2 \sum_{i=1}^{T} (1+2\eta)^{T-i} \|\mathrm{diag}(\mathbb{E}[B_{i-1}])\|. \tag{8.15}$$

Factoring out $(1+2\eta)^T$, the inequality writes

$$\mathbb{E}[V_T] \leq (1+2\eta)^T \Big(-\varepsilon w_{0,1}^2 + \frac{1}{(1+2\eta)^T} + \frac{(1+2\eta(1-\varepsilon))^T}{(1+2\eta)^T \eta(T+1)} + \eta^2 d^2 \sum_{i=1}^{T} (1+2\eta)^{-i} \|\mathrm{diag}(\mathbb{E}[B_{i-1}])\| \Big) \tag{8.16}$$

To simplify computation, we want to have a uniform bound on the spectral norm of

diag($\mathbb{E}[B_k]$). Lemma 8.6 implies that

$$\begin{aligned}
\|\mathrm{diag}(\mathbb{E}[B_k])\| &\leq 2\frac{\eta}{\eta d^2+1}\left(\frac{1}{1-\eta(\eta d^2+2)}-\frac{1}{1-\eta}\right)(1-\varepsilon) \\
&\quad + 2\frac{\eta}{\eta d^2+1}\Big(\eta d^2\frac{1}{1-\eta(\eta d^2+2)} \qquad (8.17)\\
&\qquad + \frac{1}{1-\eta}\Big)(2-\varepsilon) + \left(1+\frac{\eta^2 d^2}{1-\eta(\eta d^2+2)}\right)(1-\varepsilon) \\
&\leq 2\frac{\eta}{\eta d^2+1}\left(\frac{1-\varepsilon+(2-\varepsilon)\eta d^2}{1-\eta(\eta d^2+2)}+\frac{1}{1-\eta}\right)+\left(1+\frac{\eta^2 d^2}{1-\eta(\eta d^2+2)}\right)(1-\varepsilon) \\
&\leq 2\frac{\eta}{\eta d^2+1}\frac{2-\varepsilon+(2-\varepsilon)\eta d^2}{1-\eta(\eta d^2+2)}+1+\frac{\eta^2 d^2}{1-\eta(\eta d^2+2)} \qquad (8.18)
\end{aligned}$$

for all k. This simplifies into

$$\|\mathrm{diag}(\mathbb{E}[B_k])\| \leq 1+\frac{\eta^2 d^2+4\eta}{1-\eta(\eta d^2+2)}. \qquad (8.19)$$

Thus we obtain

$$\begin{aligned}
\mathbb{E}[V_T] &\leq (1+2\eta)^T\Big(-\varepsilon w_{0,1}^2+\frac{1}{(1+2\eta)^T}+\frac{(1+2\eta(1-\varepsilon))^T}{(1+2\eta)^T\eta(T+1)} \\
&\quad + \eta^2 d^2\left(1+\frac{\eta^2 d^2+4\eta}{1-\eta(\eta d^2+2)}\right)\sum_{i=1}^{T}(1+2\eta)^{-i}\Big) \qquad (8.20)
\end{aligned}$$

Bounding $\sum_{i=1}^{T}(1+2\eta)^{-i}$ by its infinite series $\sum_{i=1}^{\infty}(1+2\eta)^{-i}=(2\eta)^{-1}$ yields

$$\begin{aligned}
\mathbb{E}[V_T] &\leq (1+2\eta)^T\Big(-\varepsilon w_{0,1}^2+\frac{1}{(1+2\eta)^T}+\frac{(1+2\eta(1-\varepsilon))^T}{(1+2\eta)^T\eta(T+1)} \qquad (8.21)\\
&\quad + \eta/2d^2\left(1+\frac{\eta^2 d^2+4\eta}{1-\eta(\eta d^2+2)}\right)\Big). \qquad (8.22)
\end{aligned}$$

We can show that for well chosen η and T, the term under parenthesis is less that $-\varepsilon/4p$. Taking for example $\eta=\frac{\varepsilon}{4Cpd^2}$ for some constant C such that $\left(1+\frac{\eta^2 d^2+4\eta}{1-\eta(\eta d^2+2)}\right)\leq 2$ and $T>\max(4p^2d^2C/\varepsilon, \log(4p\varepsilon^{-1})/\log(1+\varepsilon/(Cpd^2)))$ gives the result. For a small enough ε, we can take $C=1$.

8.4 Implementation

8.4.1 Choosing the learning rate

In this section, we address the question of choosing the learning rate, i.e. the step-size η in iterations (8.4). Tuning the learning rate is essential in practice as it is well known to have a huge impact on the convergence speed of the method. Our idea to tune the learning rate is as follows:

- Choose the tolerance $\epsilon \in (0,1)$, and the algorithm's parameters $R,\ K \in \mathbb{N}_*$, $\rho \in (0,1)$ and $\beta > 0$.

- *Burn-in period:*

  - For $\eta \in \{\rho^k\}_{k=1:K}$, run R gradient iterations in parallel whose iterates are denoted by $\mathbf{w}_t^{(k,r)}$, $t = 1, \ldots, B$.

  - Define $\pi_0^{(k)} = 1/K$, $k = 1, \ldots, K$. For $t = 1, \ldots, B$, let

$$L_t^{(k)} = \frac{2}{R(R-2)} \sum_{r<r'=2,\ldots,R} \langle \mathbf{w}_t^{(k,r)}, \mathbf{w}_t^{(k,r')} \rangle, \tag{8.23}$$

    and for $k = 1, \ldots, K$, define

$$\pi_{t+1}^{(k)} = \pi_t^{(k)} \ \exp\left(\beta \ L_t^{(k)}\right). \tag{8.24}$$

  - Stop when $\max_{k=1,\ldots,K} \ L_t^{(k)} \geq 1 - 10\ \epsilon$.

- *After burn-in*:

  - Reset R to 1 and K to 1.

  – Normalise π.

  - At each step $t = B+1, \ldots$, choose the stepsize with probability π_B.

  - Stop when $L_t^{(1)} \geq 1 - \epsilon$.

Choosing the parameter β is more robust than choosing the learning rate. Moreover, a reasonably effective value for β is given by (see [88]):

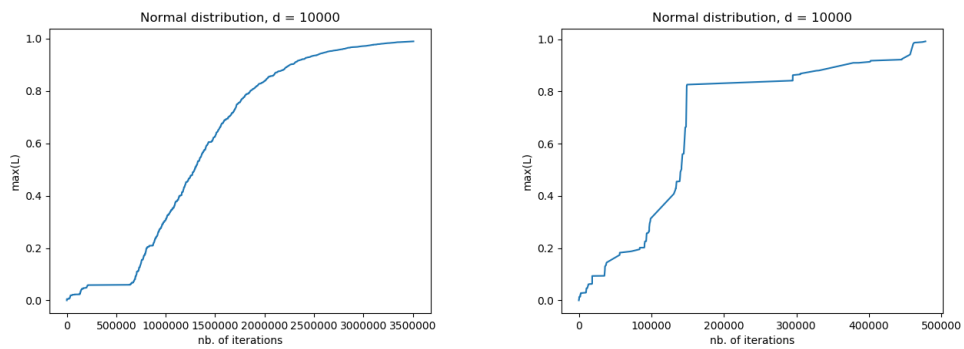
$$\beta = \sqrt{\frac{\log(K)}{B}}. \tag{8.25}$$

8.4.2 Numerical experiment

In this section, we present a simple numerical experiment which shows that

- The stochastic gradient method actually works in practice
- The adaptive selection of the learning rate/step-size described in the previous subsection actually accelerates the method's convergence drastically.

We run a simple experiment on a random i.i.d. Gaussian matrix of size 10000×10000. The convergence of $(L_t^{(1)})_{t\in\mathbb{N}}$ to 1 of the plain stochastic gradient method is shown in Figure 8.1a below. The accelerated version's convergence for the same experiment is shown in Figure 8.1b below. These results show that the method of the previous Section actually provides a substantial acceleration. We carefully checked that the selected learning rate is not equal to the smallest nor the largest value on the proposed grid of values between 2^{-3}, 2^{-2}, $\ldots 2^{17}$. The observed gain in convergence speed was by a factor of 8.75. Extensive numerical experiment demonstrating this behaviour at larger scales will be included in an expanded version of this work.



(a) $K = 1$ (b) $K = 20$

Figure 8.1: Convergence of $(L_t^{(1)})_{t\in\mathbb{N}}$ as a function of the iteration index: (a) is for the case of the arbitrary choice of learning rate equal to 2^-4 and (b) shows the behaviour of the method using the learning procedure of Section 8.4.1 for values of the learning rate equal to 2^{-3}, 2^{-2}, 2^{-1}, 1, 2, $\ldots$, 2^{17}.

8.5 Conclusion

In the present paper, we studied the convergence of the projected stochastic gradient for online principal component analysis without eigengap assumption. We showed that in expectation, the algorithm converge to a leading eigenvector. Future possible work include adapting the result in a batch setting where $k > 1$ eigenvectors are extracted at once. Otherwise, we can try to find concentration results which would help proving a convergence with high probability.

Appendices

8.A Technical lemmæ

Recall that

$$B_T = \prod_{t=T}^{1} (I + \eta A_t)^\top ((1-\varepsilon)I - A) \prod_{t=1}^{\top} (I + \eta A_t). \tag{8.26}$$

Lemma 8.2. *In the case of matrix completion, given a matrix X, we have*

$$\mathbb{E}[A_t^\top X A_t] = d^2 \ \mathrm{diag}(A \ \mathrm{diag}(X) A).$$

Proof. The resulting matrix writes

$$\begin{aligned} A_t^\top X A_t &= d^4 A_{ij} A_{ji} \mathbf{e}_{j_t} \mathbf{e}_{i_t}^\top X \mathbf{e}_{i_t} \mathbf{e}_{j_t}^\top \\ &= d^4 A_{ij} A_{ji} X_{ii} \mathbf{e}_{j_t} \mathbf{e}_{j_t}^\top. \end{aligned}$$

Therefore the expected matrix writes

$$\mathbb{E}[A_t^\top X A_t] = d^2 \sum_{i,j}^{d} A_{ij} A_{ji} X_{ii} \mathbf{e}_j \mathbf{e}_j^\top$$

Using the symmetry of A gives the result. □

Now our next goal is to see how $\mathrm{diag}\left(A^\top \mathrm{diag}(\mathbb{E}[B_{T-1}]) A\right)$ evolves with the iterations. For this purpose, take the diagonal of (8.11), multiply from the left by $A^\top$ and from the right by A and take the diagonal of the resulting expression.

Lemma 8.3. *We have that*

$$\|\mathrm{diag}\left(\mathbb{E}[B_T]\right)\| \leq 2\eta \ \|\mathbb{E}[B_{T-1}]\|_{1\to 2} + (1 + \eta^2 d^2) \ \|\mathrm{diag}(\mathbb{E}[B_{T-1}])\| \tag{8.27}$$

Proof. Expanding the recurrence relationship (8.11) gives

$$\begin{aligned} \mathrm{diag}(\mathbb{E}[B_T]) = \mathrm{diag}(\mathbb{E}[B_{T-1}]) &+ \eta \left(\mathrm{diag}\left(A^\top \mathbb{E}[B_{T-1}] + \mathbb{E}[B_{T-1}] A\right)\right) \\ &+ \eta^2 d^2 \mathrm{diag}\left(A^\top \mathrm{diag}(\mathbb{E}[B_{T-1}]) A\right). \end{aligned}$$

For any diagonal matrix Δ and symmetric matrix A, we have

$$\|\mathrm{diag}(A^\top \Delta A)\| \leq \|A\|_{1\to 2}^2 \|\Delta\|. \tag{8.28}$$

Therefore, by taking the operator norm on both sides of the equality, we have

$$\|\mathrm{diag}(\mathbb{E}[B_T])\| \leq (1+\eta^2 d^2 \|A\|_{1\to 2}^2)\|\mathrm{diag}(\mathbb{E}[B_{T-1}])\| + 2\eta \|\mathrm{diag}(A^\top \mathbb{E}[B_{T-1}])\| \tag{8.29}$$

We conclude using $\|\mathrm{diag}(A^\top E[B_{T-1}])\| \leq \|A\|_{1\to 2}\|\mathbb{E}[B_{T-1}]\|_{1\to 2}$ and $\|A\|_{1\to 2} \leq 1$. □

We also have to understand how the $\ell_{1\to 2}$ norm evolves.

Lemma 8.4. *We have*

$$\|\mathbb{E}[B_T]\|_{1\to 2} \leq \eta\ \|\mathbb{E}[B_{T-1}]\| + (1+\eta)\ \|\mathbb{E}[B_{T-1}]\|_{1\to 2} + \eta^2 d^2\ \|\mathrm{diag}(\mathbb{E}[B_{T-1}])\|. \tag{8.30}$$

Proof. Expanding the recurrence relationship gives

$$\begin{aligned}\|\mathbb{E}[B_T]\|_{1\to 2} = \|\mathbb{E}[B_{T-1}]\|_{1\to 2} + \eta\ \left(\|A^\top \mathbb{E}[B_{T-1}]\|_{1\to 2} + \|\mathbb{E}[B_{T-1}]^\top A\|_{1\to 2}\right) \\ + \eta^2 d^2 \|\mathrm{diag}(A^\top \mathrm{diag}(\mathbb{E}[B_{T-1}])A)\|_{1\to 2}.\end{aligned}$$

For a diagonal matrix Δ, we have $\|\Delta\|_{1\to 2} = \|\Delta\|$. This leads to

$$\begin{aligned}\|\mathbb{E}[B_T]\|_{1\to 2} = \|\mathbb{E}[B_{T-1}]\|_{1\to 2} + \eta\ (\|A\|\|E[B_{T-1}]\|_{1\to 2} + \|\mathbb{E}[B_{T-1}]\|\|A\|_{1\to 2}) \\ + \eta^2 d^2 \|A\|_{1\to 2}^2 \|\mathrm{diag}(\mathbb{E}[B_{T-1}])\|.\end{aligned}$$

Finally, using $\|A\|_{1\to 2} \leq 1$ concludes the proof. □

We then have to understand how the operator norm of $\mathbb{E}[B_T]$ evolves

Lemma 8.5. *We have*

$$\|\mathbb{E}[B_T]\| \leq (1+2\eta)\|\mathbb{E}[B_{T-1}]\| + \eta^2 d^2\ \|\mathrm{diag}(\mathbb{E}[B_{T-1}])\|. \tag{8.31}$$

Proof. Expanding the recurrence relationship (8.11) return

$$\|\mathbb{E}[B_T]\| = \mathbb{E}[B_{T-1}] + \eta(\|A^\top \mathbb{E}[B_{T-1}]\| + \|\mathbb{E}[B_{T-1}]A\|) + \eta^2 d^2 \|\mathrm{diag}(A^\top \mathrm{diag}(\mathbb{E}[B_{T-1}])A)\|.$$

Then using similar inequalities as in the proof of the lemmas above, we have the result. □

Lemma 8.6. *Let* $\|A\| = 1$*, then we have*

$$\|\mathrm{diag}(\mathbb{E}[B_T])\| \leq \alpha \max_j (1-\varepsilon - s_j) + \beta \|(1-\varepsilon)I - A\|_{1\to 2} + \gamma \max_j (1-\varepsilon - A_{jj}) \tag{8.32}$$

where

$$\begin{aligned}\alpha &= 2\frac{\eta}{\eta d^2+1}\left(\frac{1-\eta^{T-2}(\eta d^2+2)^{T-2}}{1-\eta(\eta d^2+2)} - \frac{1-\eta^{T-2}}{1-\eta}\right) \\ \beta &= 2\frac{\eta}{\eta d^2+1}\left(\eta d^2 \frac{1-\eta^{T-2}(\eta d^2+2)^{T-2}}{1-\eta(\eta d^2+2)} + \frac{1-\eta^{T-2}}{1-\eta}\right) \\ \gamma &= 1+\eta^2 d^2 \frac{1-\eta^{T-2}(\eta d^2+2)^{T-2}}{1-\eta(\eta d^2+2)}\end{aligned}$$

Proof. Expanding the recurrence and using equations (8.27), (8.30), and (8.31) yields the following system

$$\begin{bmatrix} \|\mathbb{E}[B_T]\| \\ \|\mathbb{E}[B_T]\|_{1\to 2} \\ \|\mathrm{diag}(\mathbb{E}[B_T])\| \end{bmatrix} \leq \left(I + \eta \begin{bmatrix} 2 & 0 & \eta d^2 \\ 1 & 1 & \eta d^2 \\ 0 & 2 & \eta d^2 \end{bmatrix} \right) \begin{bmatrix} \|\mathbb{E}[B_{T-1}]\| \\ \|\mathbb{E}[B_{T-1}]\|_{1\to 2} \\ \|\mathrm{diag}(\mathbb{E}[B_{T-1}])\| \end{bmatrix} \tag{8.33}$$

To obtain the result, we expand the inequality by recurrence. Therefore, we are interested in computing the T-th power of the matrix in inequality (8.33). We have

$$\left(I + \eta \begin{bmatrix} 2 & 0 & \eta d^2 \\ 1 & 1 & \eta d^2 \\ 0 & 2 & \eta d^2 \end{bmatrix} \right)^T = I + \sum_{i=1}^{T} \eta^i \begin{bmatrix} 2 & 0 & \eta d^2 \\ 1 & 1 & \eta d^2 \\ 0 & 2 & \eta d^2 \end{bmatrix}^i . \tag{8.34}$$

After computing the power matrices, it result that

$$\begin{aligned} \|\mathrm{diag}(\mathbb{E}[B_T])\| \leq & \sum_{i=1}^{T} \left(\eta^i \frac{2(\eta d^2 + 2)^{i-1} - 1}{\eta d^2 + 1} \right) \|\mathbb{E}[B_0]\| \\ & + \sum_{i=1}^{T} \left(\eta^i \frac{2\eta d^2 (\eta d^2 + 2)^{i-1} + 1}{\eta d^2 + 1} \right) \|\mathbb{E}[B_0]\|_{1\to 2} \\ & + \left(1 + \eta^2 d^2 \sum_{i=1}^{T} (\eta^2 d^2 + 2\eta)^{i-1} \right) \|\mathrm{diag}(\mathbb{E}[B_0])\|. \end{aligned} \tag{8.35}$$

We conclude after computing the sums and bounding from above $\|\mathbb{E}[B_0]\|$ by $\max_j (1 - \varepsilon - s_j)$. □

Lemma 8.7. *For $\eta < 1$ and $\varepsilon > 0$, we have*

$$\max_{s \in [0,1]} (1 + 2\eta\ s)^T (1 - \varepsilon - s) \leq 1 + \frac{(1 + 2\eta(1 - \varepsilon))^T}{\eta(T + 1)} \tag{8.36}$$

Proof. Denote $f(s) = (1 + 2\eta\ s)^T (1 - \varepsilon - s)$. Differentiating f and setting to zero, we obtain

$$\begin{aligned} & 2\eta T (1 + 2\eta\ s)^{T-1} (1 - \varepsilon - s) - (1 + 2\eta\ s)^T = 0 \\ & \iff 2\eta T (1 - \varepsilon - s) - (1 + 2\eta\ s) = 0 \\ & \iff \frac{T(1 - \varepsilon) - 1/2\eta}{T + 1} = s \end{aligned}$$

Let $s_c = \frac{T - \varepsilon - 1/2\eta}{T+1}$ denote this critical point. Consider the two following cases :

- if $s_c \notin [0, 1]$, then f has no critical point in the domain and therefore is maximised at either domain endpoint, i.e.

$$\max_{s \in [0,1]} f(s) = \max\{f(0) = 1 - \varepsilon, f(1) = -\varepsilon(1 + 2\eta)^T\} \leq 1$$

- if $s_c \in [0,1]$, then f is maximised at s_c and the value of f at s_c is

$$\begin{aligned}
&\left(1+2\eta\frac{T(1-\varepsilon)-1/2\eta}{T+1}\right)^T\left(1-\varepsilon-\frac{T(1-\varepsilon)-1/2\eta}{T+1}\right)\\
&=\left(1+\frac{2\eta T(1-\varepsilon)-1}{T+1}\right)^T\left(\frac{1-\varepsilon+1/2\eta}{T+1}\right)\\
&\leq (1+2\eta(1-\varepsilon))^T\left(\frac{1+1/2\eta}{T+1}\right)\\
&\leq \frac{(1+2\eta(1-\varepsilon))^T}{\eta(T+1)}.
\end{aligned}$$

Overall, the maximum value that f can reach is less than $\max\{1, \frac{(1+2\eta(1-\varepsilon))^T}{\eta(T+1)}\} \leq 1 + \frac{(1+2\eta(1-\varepsilon))^T}{\eta(T+1)}\}$. Hence the result. □

Chapter 9

Perspectives

Dans cette section, on présentera quelques ouvertures possibles aux travaux présentés.

9.1 Perspectives suivant les travaux rencontrés dans la partie I

Les différents chapitres conduisent à quelques questions et perspectives. On présentera ici trois ouvertures possibles.

La première ouverture consisterai à une étude de comparaison entre la méthode de vraisemblance exacte et la méthode de vraisemblance censurée pour le modèle Hüsler-Reiss Pareto. Plus précisément, dans le cas où le modèle est misspécifié (données dans le domaine d'attraction), Wadsworth et Tawn [180] ont montré que, dans le cas max-stable, la censure des "petites valeurs" permet de réduire le biais d'estimation. On peut alors commencer par une étude numérique pour comparer les deux méthodes pour voir si la méthode censurée est utile pour améliorer les résultats. Puis, on peut passer à une étude théorique des propriétés asymptotique de convergence et de normalité des estimateurs dans le cas censurée.

La seconde ouverture est plus appliquée. On peut exploiter la structure exponentielle du modèle Hüsler-Reiss Pareto pour modéliser la dépendance des extrêmes multivariés en présence de covariable à l'aide de modèle Vector Generalised Linear Models (VGLM) ou Vector Generalised Additive Model (VGAM). Une telle approche n'est pas sans précédent. Ainsi Chavez-Demoulin et Davison [47] ont considéré dans le cas des excès au-dessus d'un seuil univarié un modèle GAM utilisant des splines comme lisseurs. Dans le cas d'extrèmes multivariés, Carvalho et Davison [45] ainsi que Sharma, Chavez-Demoulin et Dillenbourg [155] ont considérés des modèles VGLM/VGAM. La difficultés qui semblent apparaître concernant l'usage de modèle VGLM ou VGAM sont les contraintes sur les paramètres du modèle Hüsler-Reiss Pareto qui doivent être respecté dans les modèles VGLM/GAM, en particulier, la propriété sur Q d'être semi-définie positive semble être une difficulté à surmonter. Un exemple d'application avec des données réelles serait de considérer les log-rendements de plusieurs actifs du CAC40 avec comme covariable le log-rendement de l'indice du CAC40. Les références sur les modèles VGAM/VGLM qui nous serviront dans cette approche sont

Yee [185] pour la théorie générale avec des implémentations sur R et sa mise en contexte dans la théorie des extrêmes de Yee [186].

Finalement, étant donné la structure de famille exponentielle du modèle Hüsler-Reiss Pareto, on retrouve une famille conjuguée [Proposition 3.3.13 [143]] pour des approches bayésiennes. La question naturelle est la suivante : peut-on utiliser la propriété famille exponentielle du modèle Hüsler-Reiss Pareto pour une méthodologie bayésienne multivariés? Une méthodologie bayésienne dans le cas max-stable à été proposé par Dombry, Engelke, Oesting [68]. Mais le cas des modèles des excès au-dessus d'un seuil multivarié, il n'existe pas à nos connaissance de travaux de type bayésien. La difficulté première d'une telle approche est la complexité de la famille conjuguée qui mérite d'être l'objet d'une étude plus poussée.

9.2 Perspectives suivant les travaux présentés dans la partie II

Les chapitres de la second partie conduisent aussi à quelques questions et perspectives. Certaines sont présentés à la fin de chaque chapitre mais d'autres non. On rappellera ici quelques ouvertures possibles.

Suite aux travaux du chapitre 7, la question est de savoir quand est-ce que la borne du théorème est petite? Ce qui se ramènerai à trouver des conditions sur le graphe original. Ainsi si on peut obtenir des conditions facilement vérifiable pour des grands graphes, alors notre approche serait intéressante que se soit pour le problème de sélection de sous-ensemble indépendant maximal ou encore le problème d'extraction initial (sous réserve qu'on puisse traduire la condition sur le graphe vers une condition sur la matrice de départ). Par ailleurs, ce chapitre est pour l'instant sans étude numérique, un autre projet serai donc d'illustrer numériquement l'efficacité de la méthode proposée.

Suite aux travaux du chapitre 8, une piste de recherche est l'étude d'inégalité de concentration autour de l'espérance afin d'obtenir la convergence avec forte probabilité des itérés. Trouver une telle inégalité serait idéale, néanmoins, il est possible de montrer la convergence avec forte probabilité sans avoir une telle inégalité (par exemple [151]). Par ailleurs, on avait considéré l'analyse en composante principale pour un problème de complétion matricielle. Un sujet d'ouverture possible serait de considérer le cas où l'échantillon serait en partie censurée. Par exemple, si on n'observe que deux composantes, on peut au mieux étudier la covariance entre les deux composantes (et la variance). La question est alors : est-ce qu'on converge toujours vers la bonne matrice de covariance? Sous quelle condition sur la censure?

D'autres pistes de recherche seraient des études plus poussée sur l'applicabilité des algorithmes stochastique pour les problèmes de sélection de sous-matrices. On avait proposé un algorithme greedy dans le chapitre 5 mais est-ce qu'on ne pourrait améliorer l'algorithme en prenant les colonnes aléatoirement tout en se servant de nos résultats?

Liste des publications

[53] S. Chretien, C. Guyeux, and Z. W. O. HO. "Average performance analysis of the stochastic gradient method for online PCA". In: *ArXiv e-prints (accepté à LOD2018)* (Apr. 2018). arXiv: `1804.01071 [math.ST]`.

[54] S. Chrétien and Z. W. O. Ho. "Feature Selection in Weakly Coherent Matrices". In: *Latent Variable Analysis and Signal Separation*. Ed. by Yannick Deville et al. Cham: Springer International Publishing, 2018, pp. 127–138. ISBN: 978-3-319-93764-9.

[103] Z. W. O. HO and C. Dombry. "Simple models for multivariate regular variations and the Hüsler-Reiss Pareto distribution". In: *ArXiv e-prints (soumis à JMVA (en révision))* (Dec. 2017). arXiv: `1712.09225 [stat.ME]`.

Bibliographie

[1] B. ADCOCK et al. "BREAKING THE COHERENCE BARRIER: A NEW THEORY FOR COMPRESSED SENSING". In: *Forum of Mathematics, Sigma* 5 (2017), e4. DOI: 10.1017/fms.2016.32.

[2] B. Adcock et al. *Breaking the coherence barrier: asymptotic incoherence and asymptotic sparsity in compressed sensing*. Tech. rep. 2013.

[3] C. C. Aggarwal and C. Zhai. "A Survey of Text Classification Algorithms". In: *Mining Text Data*. Ed. by Charu C. Aggarwal and ChengXiang Zhai. Boston, MA: Springer US, 2012, pp. 163–222. ISBN: 978-1-4614-3223-4. DOI: 10.1007/978-1-4614-3223-4_6. URL: https://doi.org/10.1007/978-1-4614-3223-4_6.

[4] Z. Allen-Zhu and Y. Li. "LazySVD: Even faster SVD decomposition yet without agonizing pain". In: *Advances in Neural Information Processing Systems*. 2016, pp. 974–982.

[5] S. Arora, R. Ge, and A. Moitra. "New algorithms for learning incoherent and overcomplete dictionaries". In: *Conference on Learning Theory*. 2014, pp. 779–806.

[6] J.-M. Azaïs, S. Mourareau, and Y. De Castro. "A rice method proof of the null-space property over the Grassmannian". In: *Ann. Inst. H. Poincaré Probab. Statist.* 53.4 (Nov. 2017), pp. 1821–1838. DOI: 10.1214/16-AIHP772. URL: https://doi.org/10.1214/16-AIHP772.

[7] A. A. Balkema and L. de Haan. "Residual life time at great age". In: *Ann. Probability* 2 (1974), pp. 792–804.

[8] A. A. Balkema and S. I. Resnick. "Max-infinite divisibility". In: *J. Appl. Probability* 14.2 (1977), pp. 309–319. ISSN: 0021-9002.

[9] A. S. Bandeira. *Ten Lectures and Forty-Two Open Problems in the Mathematics of Data Science*. 2015.

[10] A. S. Bandeira et al. "The road to deterministic matrices with the restricted isometry property". In: *Journal of Fourier Analysis and Applications* 19.6 (2013), pp. 1123–1149.

[11] R. G. Baraniuk. "Compressive sensing [lecture notes]". In: *IEEE signal processing magazine* 24.4 (2007), pp. 118–121.

[12] R. Baraniuk et al. "A Simple Proof of the Restricted Isometry Property for Random Matrices". In: *Constructive Approximation* 28.3 (Dec. 2008), pp. 253–263. ISSN: 1432-0940. DOI: 10.1007/s00365-007-9003-x. URL: https://doi.org/10.1007/s00365-007-9003-x.

[13] A. Barg, A. Mazumdar, and R. Wang. "Restricted isometry property of random subdictionaries". In: *Information Theory, IEEE Transactions on* 61.8 (2015), pp. 4440–4450.

[14] O. Barndorff-Nielsen. *Information and exponential families in statistical theory*. Wiley Series in Probability and Statistics. Reprint of the 1978 original [MR0489333]. John Wiley & Sons, Ltd., Chichester, 2014, pp. x+238. ISBN: 978-1-118-85750-2. DOI: 10.1002/9781118857281. URL: http://dx.doi.org/10.1002/9781118857281.

[15] V. Barnett. "The ordering of multivariate data". In: *J. Roy. Statist. Soc. Ser. A* 139.3 (1976). With a discussion by R. L. Plackett, K. V. Mardia, R. M. Loynes, A. Huitson, G. M. Paddle, T. Lewis, G. A. Barnard, A. M. Walker, F. Downton, P. J. Green, Maurice Kendall, A. Robinson, Allan Seheult and D. H. Young, pp. 318–355. ISSN: 0035-9238. DOI: 10.2307/2344839. URL: https://doi.org/10.2307/2344839.

[16] J. Batson, D. A. Spielman, and N. Srivastava. "Twice-ramanujan sparsifiers". In: *SIAM Journal on Computing* 41.6 (2012), pp. 1704–1721.

[17] J. Batson et al. "Spectral sparsification of graphs: theory and algorithms". In: *Communications of the ACM* 56.8 (2013), pp. 87–94.

[18] J. Beirlant et al. *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2004. ISBN: 9780471976479. URL: https://books.google.fr/books?id=GtIYLAlTcKEC.

[19] P. C. Bellec. "Localized Gaussian width of M-convex hulls with applications to Lasso and convex aggregation". In: *arXiv preprint arXiv:1705.10696* (2017).

[20] L. R. Belzile and J. G. Nešlehová. "Extremal attractors of Liouville copulas". In: *Journal of Multivariate Analysis* 160 (2017), pp. 68–92. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2017.05.008. URL: http://www.sciencedirect.com/science/article/pii/S0047259X17300453.

[21] R. B. Bendel and A. A. Afifi. "Comparison of Stopping Rules in Forward "Stepwise" Regression". In: *Journal of the American Statistical Association* 72.357 (1977), pp. 46–53. ISSN: 01621459. URL: http://www.jstor.org/stable/2286904.

[22] A. Ben-Hur and I. Guyon. "Detecting stable clusters using principal component analysis". In: *Functional genomics*. Springer, 2003, pp. 159–182.

[23] J. Bennett, S. Lanning, et al. "The netflix prize". In: *Proceedings of KDD cup and workshop*. Vol. 2007. New York, NY, USA. 2007, p. 35.

[24] R. Bhatia. *Perturbation bounds for matrix eigenvalues*. SIAM, 2007.

[25] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[26] J. Bourgain and L. Tzafriri. "Invertibility of 'large' submatrices with applications to the geometry of Banach spaces and harmonic analysis". In: *Israel journal of mathematics* 57.2 (1987), pp. 137–224.

[27] C. Boutsidis, P. Drineas, and M. Magdon-Ismail. "Near-optimal column-based matrix reconstruction". In: *SIAM Journal on Computing* 43.2 (2014), pp. 687–717.

[28] C. Boutsidis, M. W. Mahoney, and P. Drineas. "An improved approximation algorithm for the column subset selection problem". In: *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics. 2009, pp. 968–977.

[29] L. Breiman. "On some limit theorems similar to the arc-sin law". In: *Theory of Probability and its Applications* 10.2 (1965), pp. 323–331.

[30] B. M. Brown and S. I. Resnick. "Extreme values of independent stochastic processes". In: *J. Appl. Probability* 14.4 (1977), pp. 732–739. ISSN: 0021-9002.

[31] J. Bruna and S. Mallat. "Invariant scattering convolution networks". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1872–1886.

[32] P. Bühlmann and S. A. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[33] J. Cahill, X. Chen, and R. Wang. "The gap between the null space property and the restricted isometry property". In: *Linear Algebra and its Applications* 501 (2016), pp. 363–375. ISSN: 0024-3795. DOI: https://doi.org/10.1016/j.laa.2016.03.022. URL: http://www.sciencedirect.com/science/article/pii/S0024379516300088.

[34] E. J. Candes. "Mathematics of sparsity (and a few other things)". In: *Proceedings of the International Congress of Mathematicians, Seoul, South Korea*. Vol. 123. Citeseer. 2014.

[35] E. J. Candes. "The restricted isometry property and its implications for compressed sensing". In: *Comptes Rendus Mathematique* 346.9 (2008), pp. 589–592.

[36] E. J. Candes and J. Romberg. "Sparsity and incoherence in compressive sampling". In: *Inverse problems* 23.3 (2007), p. 969.

[37] E. J. Candes, J. Romberg, and T. Tao. "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on Information Theory* 52.2 (Feb. 2006), pp. 489–509. ISSN: 0018-9448. DOI: 10.1109/TIT.2005.862083.

[38] E. J. Candes, K. J. Romberg, and T. Tao. "Stable Signal Recovery from Incomplete and Inaccurate Measurements". In: 59 (Aug. 2006).

[39] E. J. Candes and T. Tao. "Decoding by linear programming". In: *IEEE Transactions on Information Theory* 51.12 (Dec. 2005), pp. 4203–4215. ISSN: 0018-9448. DOI: 10.1109/TIT.2005.858979.

[40] E. J. Candes and M. B Wakin. "An introduction to compressive sampling". In: *IEEE signal processing magazine* 25.2 (2008), pp. 21–30.

[41] E. J. Candes et al. "Compressed sensing with coherent and redundant dictionaries". In: *Applied and Computational Harmonic Analysis* 31.1 (2011), pp. 59–73.

[42] E. J. Candes, Y. Plan, et al. "Near-ideal model selection by ℓ_1 minimization". In: *The Annals of Statistics* 37.5A (2009), pp. 2145–2177.

[43] E. J. Candes et al. "Robust principal component analysis?" In: *Journal of the ACM (JACM)* 58.3 (2011), p. 11.

[44] H. Cardot and D. Degras. "Online Principal Component Analysis in High Dimension: Which Algorithm to Choose?" In: *arXiv preprint arXiv:1511.03688* (2015).

[45] M. de Carvalho and A. C. Davison. "Spectral Density Ratio Models for Multivariate Extremes". In: *Journal of the American Statistical Association* 109.506 (2014), pp. 764–776. DOI: 10.1080/01621459.2013.872651. eprint: https://doi.org/10.1080/01621459.2013.872651. URL: https://doi.org/10.1080/01621459.2013.872651.

[46] V. Cevher et al. "Near-optimal Bayesian localization via incoherence and sparsity". In: *Information Processing in Sensor Networks, 2009. IPSN 2009. International Conference on*. IEEE. 2009, pp. 205–216.

[47] V. Chavez-Demoulin and A. C. Davison. "Generalized additive modelling of sample extremes". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54.1 (), pp. 207–222. DOI: 10.1111/j.1467-9876.2005.00479.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9876.2005.00479.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2005.00479.x.

[48] S. Chen, D. Donoho, and M. Saunders. "Atomic Decomposition by Basis Pursuit". In: *SIAM Journal on Scientific Computing* 20.1 (1998), pp. 33–61. DOI: 10.1137/S1064827596304010. eprint: https://doi.org/10.1137/S1064827596304010. URL: https://doi.org/10.1137/S1064827596304010.

[49] S. Chretien and S. Darses. "An elementary approach to the problem of column selection in a rectangular matrix". In: *arXiv preprint arXiv:1509.00748* (2015).

[50] S. Chretien and S. Darses. "Perturbation bounds on the extremal singular values of a matrix after appending a column". In: *arXiv preprint arXiv:1406.5441* (2014).

[51] S. Chretien and S. Darses. "Sparse recovery with unknown variance: a LASSO-type approach". In: *Information Theory, IEEE Transactions on* 60.7 (2014), pp. 3970–3988.

[52] S. Chrétien and S. Darses. "Invertibility of random submatrices via tail-decoupling and a matrix Chernoff inequality". In: *Statistics & Probability Letters* 82.7 (2012), pp. 1479–1487.

[53] S. Chretien, C. Guyeux, and Z. W. O. HO. "Average performance analysis of the stochastic gradient method for online PCA". In: *ArXiv e-prints (accepté à LOD2018)* (Apr. 2018). arXiv: 1804.01071 [math.ST].

[54] S. Chrétien and Z. W. O. Ho. "Feature Selection in Weakly Coherent Matrices". In: *Latent Variable Analysis and Signal Separation*. Ed. by Yannick Deville et al. Cham: Springer International Publishing, 2018, pp. 127–138. ISBN: 978-3-319-93764-9.

[55] K. L. Chung. "On a Stochastic Approximation Method". In: *Ann. Math. Statist.* 25.3 (Sept. 1954), pp. 463–483. DOI: 10.1214/aoms/1177728716. URL: https://doi.org/10.1214/aoms/1177728716.

[56] A. Cohen, W. Dahmen, and R. DeVore. "Compressed sensing and best k-term approximation". In: *Journal of the American mathematical society* 22.1 (2009), pp. 211–231.

[57] S. Coles. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag London, Ltd., London, 2001, pp. xiv+208. ISBN: 1-85233-459-2. URL: https://doi.org/10.1007/978-1-4471-3675-0.

[58] S. G. Coles and J. A. Tawn. "Modelling extreme multivariate events". In: *J. Roy. Statist. Soc. Ser. B* 53.2 (1991), pp. 377–392. ISSN: 0035-9246. URL: http://links.jstor.org/sici?sici=0035-9246(1991)53:2%3C377:MEME%3E2.0.CO;2-4&origin=MSN.

[59] O. Dalal and B. Rajaratnam. "G-AMA: Sparse Gaussian graphical model estimation via alternating minimization". In: *ArXiv e-prints* (May 2014). arXiv: 1405.3034 [stat.CO].

[60] A. d'Aspremont and L. El Ghaoui. "Testing the nullspace property using semidefinite programming". In: *Mathematical Programming* 127.1 (Mar. 2011), pp. 123–144. ISSN: 1436-4646. DOI: 10.1007/s10107-010-0416-0. URL: https://doi.org/10.1007/s10107-010-0416-0.

[61] R. A. Davis and T. Mikosch. "Extreme value theory for space-time processes with heavy-tailed distributions". In: *Stochastic Process. Appl.* 118.4 (2008), pp. 560–584. ISSN: 0304-4149. DOI: 10.1016/j.spa.2007.06.001. URL: http://dx.doi.org/10.1016/j.spa.2007.06.001.

[62] C. De Sa, C. Ré, and K. Olukotun. "Global Convergence of Stochastic Gradient Descent for Some Non-convex Matrix Problems". In: *ICML*. 2015.

[63] P. Deheuvels. "Caractérisation complête des lois extrêmes multivariées et de la convergence aux types extrêmes". In: *Publ. Inst. Statist. Univ. Paris* 23 (1978), pp. 1–36.

[64] P. Deheuvels. "Probabilistic aspects of multivariate extremes". In: *Statistical extremes and applications (Vimeiro, 1983)*. Vol. 131. NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. Reidel, Dordrecht, 1984, pp. 117–130.

[65] J. Dennis and R. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial and Applied Mathematics, 1996. DOI: 10.1137/1.9781611971200. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611971200. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611971200.

[66] A. Deshpande and L. Rademacher. "Efficient volume sampling for row/column subset selection". In: *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*. IEEE. 2010, pp. 329–338.

[67] C. Dombry, S. Engelke, and M. Oesting. "Asymptotic properties of the maximum likelihood estimator for multivariate extreme value distributions". In: *ArXiv e-prints* (Dec. 2016). arXiv: 1612.05178 [math.ST].

[68] C. Dombry, S. Engelke, and M. Oesting. "Bayesian inference for multivariate extreme value distributions". In: *Electron. J. Statist.* 11.2 (2017), pp. 4813–4844. DOI: 10.1214/17-EJS1367. URL: https://doi.org/10.1214/17-EJS1367.

[69] C. Dombry, F. Éyi-Minko, and M. Ribatet. "Conditional simulation of max-stable processes". In: *Biometrika* 100.1 (2013), pp. 111–124. ISSN: 0006-3444. DOI: 10.1093/biomet/ass067. URL: http://dx.doi.org/10.1093/biomet/ass067.

[70] C. Dombry and A. Ferreira. "Maximum likelihood estimators based on the block maxima method". In: *ArXiv e-prints* (May 2017). arXiv: 1705.00465 [math.ST].

[71] D. L. Donoho. "Compressed sensing". In: *IEEE Transactions on Information Theory* 52.4 (Apr. 2006), pp. 1289–1306. ISSN: 0018-9448. DOI: 10.1109/TIT.2006.871582.

[72] D. L. Donoho. "Compressed sensing". In: *Information Theory, IEEE Transactions on* 52.4 (2006), pp. 1289–1306.

[73] D. L. Donoho and M. Elad. "Optimally Sparse Representation in General (Nonorthogonal) Dictionaries via ℓ_1 Minimization". In: *Proceedings of the National Academy of Sciences of the United States of America* 100.5 (2003), pp. 2197–2202. ISSN: 00278424. URL: http://www.jstor.org/stable/3139504.

[74] D. L. Donoho, M. Elad, and V. N. Temlyakov. "Stable recovery of sparse overcomplete representations in the presence of noise". In: *IEEE Transactions on Information Theory* 52.1 (Jan. 2006), pp. 6–18. ISSN: 0018-9448. DOI: 10.1109/TIT.2005.860430.

[75] D.L. Donoho. "For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution". In: *Communications on Pure and Applied Mathematics* 59.7 (), pp. 907–934. DOI: 10.1002/cpa.20131. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.20131. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20131.

[76] H. Drees, A. Ferreira, and L. de Haan. "On maximum likelihood estimation of the extreme value index". In: *Ann. Appl. Probab.* 14.3 (Aug. 2004), pp. 1179–1201. DOI: 10.1214/105051604000000279. URL: https://doi.org/10.1214/105051604000000279.

[77] M. Duflo. *Algorithmes stochastiques*. Mathématiques et Applications. Springer Berlin Heidelberg, 1996. ISBN: 9783540606994. URL: https://books.google.fr/books?id=ffkzAAAACAAJ.

[78] M. Dwass. "Extremal processes". In: *Ann. Math. Statist* 35 (1964), pp. 1718–1725. ISSN: 0003-4851. DOI: 10.1214/aoms/1177700394. URL: https://doi.org/10.1214/aoms/1177700394.

[79] B. Efron et al. "Least angle regression". In: *Ann. Statist.* 32.2 (Apr. 2004), pp. 407–499. DOI: 10.1214/009053604000000067. URL: https://doi.org/10.1214/009053604000000067.

[80] Y. C. Eldar and G. Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[81] A. K. Farahat, A. Ghodsi, and M. S. Kamel. "An efficient greedy method for unsupervised feature selection". In: *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE. 2011, pp. 161–170.

[82] A. K. Farahat, A. Ghodsi, and M. S. Kamel. "Efficient greedy feature selection for unsupervised learning". In: *Knowledge and information systems* 35.2 (2013), pp. 285–310.

[83] S. Fernández, A. Graves, and J. Schmidhuber. "Sequence Labelling in Structured Domains with Hierarchical Recurrent Neural Networks". In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. IJCAI'07. Hyderabad, India: Morgan Kaufmann Publishers Inc., 2007, pp. 774–779. URL: http://dl.acm.org/citation.cfm?id=1625275.1625400.

[84] A. Ferreira and L. de Haan. "On the block maxima method in extreme value theory: PWM estimators". In: *Ann. Statist.* 43.1 (Feb. 2015), pp. 276–298. DOI: 10.1214/14-AOS1280. URL: https://doi.org/10.1214/14-AOS1280.

[85] R. A. Fisher and L. H. C. Tippett. "Limiting forms of the frequency distribution of the largest or smallest member of a sample". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 24.2 (1928), pp. 180–190. DOI: 10.1017/S0305004100015681.

[86] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Vol. 1. 3. Birkhäuser Basel, 2013.

[87] S. Foucart and H. Rauhut. "Recovery of Random Signals using Deterministic Matrices". In: *A Mathematical Introduction to Compressive Sensing*. Springer, 2013, pp. 459–473.

[88] Y. Freund and R. E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139.

[89] Laurent Gardes and Stéphane Girard. "Asymptotic properties of a Pickands type estimator of the extreme value index". In: *Focus on Probability Theory*. Nova Science. Louis R. Velle, 2006, pp. 133–149. URL: https://hal.archives-ouvertes.fr/hal-00383148.

[90] J. Geffroy. "Contribution à la théorie des valeurs extrêmes". In: *Publ. Inst. Statist. Univ. Paris* 7.3/4 (1958), pp. 37–121.

[91] B. Gnedenko. "Sur la distribution limite du terme maximum d'une série aléatoire". In: *Ann. of Math. (2)* 44 (1943), pp. 423–453. ISSN: 0003-486X.

[92] A. Graves, A.-r. Mohamed, and G. E. Hinton. "Recognition with Deep Recurrent Neural Networks". In: 2013.

[93] R. Gribonval and M. Nielsen. "Sparse representations in unions of bases". In: *IEEE Transactions on Information Theory* 49.12 (Dec. 2003), pp. 3320–3325. ISSN: 0018-9448. DOI: 10.1109/TIT.2003.820031.

[94] G. Gudendorf and J. Segers. "Extreme-value copulas". In: *Copula theory and its applications*. Vol. 198. Lect. Notes Stat. Proc. Springer, Heidelberg, 2010, pp. 127–145. DOI: 10.1007/978-3-642-12465-5_6. URL: http://dx.doi.org/10.1007/978-3-642-12465-5_6.

[95] E. J. Gumbel. "Distributions des valeurs extrêmes en plusieurs dimensions". In: *Publ. Inst. Statist. Univ. Paris* 9 (1960), pp. 171–173.

[96] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[97] L. de Haan. *On regular variation and its application to the weak convergence of sample extremes*. Vol. 32. Mathematical Centre Tracts. Mathematisch Centrum, Amsterdam, 1970, v+124 pp. (loose errata).

[98] L. de Haan and S. I. Resnick. "Limit theory for multivariate sample extremes". In: *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 40.4 (1977), pp. 317–337. DOI: 10.1007/BF00533086. URL: http://dx.doi.org/10.1007/BF00533086.

[99] W. W. Hager. "Minimizing a quadratic over a sphere". In: *SIAM Journal on Optimization* 12.1 (2001), pp. 188–208.

[100] T. Hastie et al. "Forward stagewise regression and the monotone lasso". In: *Electronic Journal of Statistics* 1 (2007), p. 2007.

[101] E. Hazan et al. "Introduction to online convex optimization". In: *Foundations and Trends® in Optimization* 2.3-4 (2016), pp. 157–325.

[102] H.J. van den Herik, J. W.H.M. Uiterwijk, and J. van Rijswijck. "Games solved: Now and in the future". In: *Artificial Intelligence* 134.1 (2002), pp. 277–311. ISSN: 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(01)00152-7. URL: http://www.sciencedirect.com/science/article/pii/S0004370201001527.

[103] Z. W. O. HO and C. Dombry. "Simple models for multivariate regular variations and the Hüsler-Reiss Pareto distribution". In: *ArXiv e-prints (soumis à JMVA (en révision))* (Dec. 2017). arXiv: 1712.09225 [stat.ME].

[104] J. R. M. Hosking, J. R. Wallis, and E. F. Wood. "Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments". In: *Technometrics* 27.3 (1985), pp. 251–261. ISSN: 00401706. URL: http://www.jstor.org/stable/1269706.

[105] X. Huang. *Statistics of Bivariate Extreme Values: Statistiek Van Bivariate Extreme Waarden*. 1992. URL: https://books.google.fr/books?id=ewBhngEACAAJ.

[106] M. Hügel, H. Rauhut, and T. Strohmer. "Remote sensing via ℓ_1-minimization". In: *Foundations of Computational Mathematics* 14.1 (2014), pp. 115–150.

[107] H. Hult and F. Lindskog. "Regular variation for measures on metric spaces". In: *Publ. Inst. Math. (Beograd) (N.S.)* 80(94) (2006), pp. 121–140. ISSN: 0350-1302. DOI: 10.2298/PIM0694121H. URL: http://dx.doi.org/10.2298/PIM0694121H.

[108] R. Huser and A. C. Davison. "Composite likelihood estimation for the Brown–Resnick process". In: *Biometrika* 100.2 (2013), pp. 511–518. DOI: 10.1093/biomet/ass089. eprint: /oup/backfile/content_public/journal/biomet/100/2/10.1093/biomet/ass089/2/ass089.pdf. URL: +%20http://dx.doi.org/10.1093/biomet/ass089.

[109] J. Hüsler and R.-D. Reiss. "Maxima of normal random vectors: between independence and complete dependence". In: *Statist. Probab. Lett.* 7.4 (1989), pp. 283–286. ISSN: 0167-7152. DOI: 10.1016/0167-7152(89)90106-5. URL: https://doi.org/10.1016/0167-7152(89)90106-5.

[110] J. Hüsler and R.-D. Reiss. "Maxima of normal random vectors: between independence and complete dependence". In: *Statist. Probab. Lett.* 7.4 (1989), pp. 283–286. ISSN: 0167-7152. DOI: 10.1016/0167-7152(89)90106-5. URL: http://dx.doi.org/10.1016/0167-7152(89)90106-5.

[111] S.-G. Hwang. "Cauchy's Interlace Theorem for Eigenvalues of Hermitian Matrices". In: *The American Mathematical Monthly* 111.2 (2004), pp. 157–159. ISSN: 00029890, 19300972. URL: http://www.jstor.org/stable/4145217.

[112] A. F. Jenkinson. "The frequency distribution of the annual maximum (or minimum) values of meteorological elements". In: *Quarterly Journal of the Royal Meteorological Society* 81 (Apr. 1955), pp. 158–171. DOI: 10.1002/qj.49708134804.

[113] C. Jin et al. "Robust shift-and-invert preconditioning: Faster and more sample efficient algorithms for eigenvector computation". In: *arXiv preprint arXiv:1510.08896* (2015).

[114] H. Joe. *Dependence modeling with copulas*. Vol. 134. Monographs on Statistics and Applied Probability. CRC Press, Boca Raton, FL, 2015, pp. xviii+462. ISBN: 978-1-4665-8322-1.

[115] R. Johnson and T. Zhang. "Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 315–323. URL: http://dl.acm.org/citation.cfm?id=2999611.2999647.

[116] A. Kiriliouk et al. "Peaks Over Thresholds Modeling With Multivariate Generalized Pareto Distributions". In: *Technometrics* 0.0 (2018), pp. 1–13. DOI: 10.1080/00401706.2018.1462738. eprint: https://doi.org/10.1080/00401706.2018.1462738. URL: https://doi.org/10.1080/00401706.2018.1462738.

[117] P. Krupskii et al. "Extreme-value limit of the convolution of exponential and multivariate normal distributions: Link to the Hüsler–Reiß distribution". In: *Journal of Multivariate Analysis* 163 (2018), pp. 80–95. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2017.10.006. URL: http://www.sciencedirect.com/science/article/pii/S0047259X17301525.

[118] W.J. Krzanowski. "Selection of variables to preserve multivariate data structure, using principal components". In: *Applied Statistics* (1987), pp. 22–33.

[119] J. Lamperti. "On extreme order statistics". In: *Ann. Math. Statist* 35 (1964), pp. 1726–1737. ISSN: 0003-4851. DOI: 10.1214/aoms/1177700395. URL: https://doi.org/10.1214/aoms/1177700395.

[120] E. L. Lehmann. *Elements of large-sample theory*. Springer Texts in Statistics. Springer-Verlag, New York, 1999, pp. xii+631. ISBN: 0-387-98595-6. DOI: 10.1007/b98855. URL: http://dx.doi.org/10.1007/b98855.

[121] H. Luo and R. E. Schapire. "Achieving all with no parameters: Adanormalhedge". In: *Conference on Learning Theory*. 2015, pp. 1286–1304.

[122] M. W. Mahoney and P. Drineas. "CUR matrix decompositions for improved data analysis". In: *Proceedings of the National Academy of Sciences* 106.3 (2009), pp. 697–702.

[123] S. Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.

[124] S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. 3rd. Orlando, FL, USA: Academic Press, Inc., 2008. ISBN: 9780123743701.

[125] S. Mallat. "Group invariant scattering". In: *Communications on Pure and Applied Mathematics* 65.10 (2012), pp. 1331–1398.

[126] A. W. Marshall and I. Olkin. "Domains of attraction of multivariate extreme value distributions". In: *Ann. Probab.* 11.1 (1983), pp. 168–177. ISSN: 0091-1798. URL: http://links.jstor.org/sici?sici=0091-1798(198302)11:1%3C168:DOAOME%3E2.0.CO;2-S&origin=MSN.

[127] B. Nadler. "Finite sample approximation results for principal component analysis: A matrix perturbation approach". In: *The Annals of Statistics* (2008), pp. 2791–2817.

[128] E. Ndiaye et al. "Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression". In: *Journal of Physics: Conference Series* 904.1 (2017), p. 012006. URL: http://iopscience.iop.org/article/10.1088/1742-6596/904/1/012006/pdf.

[129] D. Needell, R. Zhao, and A. Zouzias. "Randomized block Kaczmarz method with projection for solving least squares". In: *Linear Algebra and its Applications* 484 (2015), pp. 322–343.

[130] J.L. Nelson and V. N. Temlyakov. "On the size of incoherent systems". In: *Journal of Approximation Theory* 163.9 (2011), pp. 1238–1245.

[131] A. Nemirovski et al. "Robust Stochastic Approximation Approach to Stochastic Programming". In: *SIAM Journal on Optimization* 19.4 (2009), pp. 1574–1609. DOI: 10.1137/070704277. eprint: https://doi.org/10.1137/070704277. URL: https://doi.org/10.1137/070704277.

[132] J. W. Neuberger. "The Continuous Newton's Method, Inverse Functions, and Nash-Moser". In: *The American Mathematical Monthly* 114.5 (2007), pp. 432–437. ISSN: 00029890, 19300972. URL: http://www.jstor.org/stable/27642222.

[133] A. K. Nikoloulopoulos, H. Joe, and H. Li. "Extreme value properties of multivariate t copulas". In: *Extremes* 12.2 (2009), pp. 129–148.

[134] H. Nyquist. "Certain Topics in Telegraph Transmission Theory". In: *Transactions of the American Institute of Electrical Engineers* 47.2 (Apr. 1928), pp. 617–644. ISSN: 0096-3860. DOI: 10.1109/T-AIEE.1928.5055024.

[135] A. B. Owen. "A robust hybrid of lasso and ridge regression". In: (2007).

[136] A. Ozçelikkale, S. Yuksel, and H. M. Ozaktas. "Unitary precoding and basis dependency of MMSE performance for Gaussian erasure channels". In: *Information Theory, IEEE Transactions on* 60.11 (2014), pp. 7186–7203.

[137] J. Pickands III. "Statistical inference using extreme order statistics". In: *Ann. Statist.* 3 (1975), pp. 119–131. ISSN: 0090-5364. URL: http://links.jstor.org/sici?sici=0090-5364(197501)3:1%3C119:SIUEOS%3E2.0.CO;2-O&origin=MSN.

[138] M. Porfiri and M. Di Bernardo. "Criteria for global pinning-controllability of complex networks". In: *Automatica* 44.12 (2008), pp. 3100–3106.

[139] B. Recht. "A simpler approach to matrix completion". In: *Journal of Machine Learning Research* 12.Dec (2011), pp. 3413–3430.

[140] S. I. Resnick. *Extreme values, regular variation, and point processes*. Vol. 4. Applied Probability. A Series of the Applied Probability Trust. Springer-Verlag, New York, 1987, pp. xii+320. ISBN: 0-387-96481-9. DOI: 10.1007/978-0-387-75953-1. URL: http://dx.doi.org/10.1007/978-0-387-75953-1.

[141] S. I. Resnick. *Heavy-tail phenomena*. Springer Series in Operations Research and Financial Engineering. Probabilistic and statistical modeling. Springer, New York, 2007, pp. xx+404. ISBN: 0-387-24272-4.

[142] H. Robbins and S. Monro. "A Stochastic Approximation Method". In: *Ann. Math. Statist.* 22.3 (Sept. 1951), pp. 400–407. DOI: 10.1214/aoms/1177729586. URL: https://doi.org/10.1214/aoms/1177729586.

[143] C. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

[144] R.T. Rockafellar. *Convex Analysis*. Princeton landmarks in mathematics and physics. Princeton University Press, 1970. ISBN: 9780691015866. URL: https://books.google.fr/books?id=1TiOka9bx3sC.

[145] J. Romberg. "Imaging via compressive sampling". In: *IEEE Signal Processing Magazine* 25.2 (2008), pp. 14–20.

[146] H. Rootzén, J. Segers, and J. L. Wadsworth. "Multivariate generalized Pareto distributions: parametrizations, representations, and properties". In: *J. Multivariate Anal.* 165 (2018), pp. 117–131. ISSN: 0047-259X. DOI: 10.1016/j.jmva.2017.12.003. URL: https://doi.org/10.1016/j.jmva.2017.12.003.

[147] H. Rootzén, J. Segers, and J. L. Wadsworth. "Multivariate peaks over thresholds models". In: *Extremes* 21.1 (Mar. 2018), pp. 115–145. ISSN: 1572-915X. DOI: 10.1007/s10687-017-0294-4. URL: https://doi.org/10.1007/s10687-017-0294-4.

[148] H. Rootzén and N. Tajvidi. "Multivariate generalized Pareto distributions". In: *Bernoulli* 12.5 (2006), pp. 917–930. ISSN: 1350-7265. DOI: 10.3150/bj/1161614952. URL: http://dx.doi.org/10.3150/bj/1161614952.

[149] J. Sacks. "Asymptotic Distribution of Stochastic Approximation Procedures". In: *Ann. Math. Statist.* 29.2 (June 1958), pp. 373–405. DOI: 10.1214/aoms/1177706619. URL: https://doi.org/10.1214/aoms/1177706619.

[150] S. Shalev-Shwartz et al. "Online learning and online convex optimization". In: *Foundations and Trends® in Machine Learning* 4.2 (2012), pp. 107–194.

[151] O. Shamir. "A Stochastic PCA and SVD Algorithm with an Exponential Convergence Rate." In: *ICML*. 2015, pp. 144–152.

[152] O. Shamir. "Convergence of Stochastic Gradient Descent for PCA". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 257–265. URL: http://dl.acm.org/citation.cfm?id=3045390.3045419.

[153] O. Shamir. "Fast Stochastic Algorithms for SVD and PCA: Convergence Properties and Convexity". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 248–256. URL: http://dl.acm.org/citation.cfm?id=3045390.3045418.

[154] C. E. Shannon. "Communication in the Presence of Noise". In: *Proceedings of the IRE* 37.1 (Jan. 1949), pp. 10–21. ISSN: 0096-8390. DOI: 10.1109/JRPROC.1949.232969.

[155] K. Sharma, V. Chavez-Demoulin, and P. Dillenbourg. "Nonstationary modelling of tail dependence of two subjects' concentration". In: *Ann. Appl. Stat.* 12.2 (June 2018), pp. 1293–1311. DOI: 10.1214/17-AOAS1111. URL: https://doi.org/10.1214/17-AOAS1111.

[156] M. Sibuya. "Bivariate extreme statistics. I". In: *Ann. Inst. Statist. Math. Tokyo* 11 (1960), pp. 195–210. ISSN: 0020-3157.

[157] D. Silver et al. "Mastering the Game of Go with Deep Neural Networks and Tree Search". In: *Nature* 529.7587 (Jan. 2016), pp. 484–489. DOI: 10.1038/nature16961.

[158] M. Sklar. "Fonctions de répartition à n dimensions et leurs marges". In: *Publ. Inst. Statist. Univ. Paris* 8 (1959), pp. 229–231.

[159] R. L. Smith, J. A. Tawn, and H. K. Yuen. "Statistics of Multivariate Extremes". In: *International Statistical Review / Revue Internationale de Statistique* 58.1 (1990), pp. 47–58. ISSN: 03067734, 17515823. URL: http://www.jstor.org/stable/1403473.

[160] D. A. Spielman and N. Srivastava. "An elementary proof of the restricted invertibility theorem". In: *Israel Journal of Mathematics* 190.1 (2012), pp. 83–91.

[161] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for machine learning*. Mit Press, 2012.

[162] V. Strassen. "The Existence of Probability Measures with Given Marginals". In: *Ann. Math. Statist.* 36.2 (Apr. 1965), pp. 423–439. DOI: 10.1214/aoms/1177700153. URL: https://doi.org/10.1214/aoms/1177700153.

[163] T. Sun and C.-H. Zhang. "Scaled sparse linear regression". In: *Biometrika* 99.4 (2012), pp. 879–898. DOI: 10.1093/biomet/ass043. eprint: /oup/backfile/content_public/journal/biomet/99/4/10.1093/biomet/ass043/2/ass043.pdf. URL: http://dx.doi.org/10.1093/biomet/ass043.

[164] J. A. Tawn. "Bivariate extreme value theory: models and estimation". In: *Biometrika* 75.3 (1988), pp. 397–415. ISSN: 0006-3444. DOI: 10.1093/biomet/75.3.397. URL: https://doi.org/10.1093/biomet/75.3.397.

[165] J. A. Tawn. "Modelling Multivariate Extreme Value Distributions". In: *Biometrika* 77.2 (1990), pp. 245–253. ISSN: 00063444. URL: http://www.jstor.org/stable/2336802.

[166] Jonathan A. Tawn. "Estimating Probabilities of Extreme Sea-Levels". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41.1 (1992), pp. 77–93. ISSN: 00359254, 14679876. URL: http://www.jstor.org/stable/2347619.

[167] J. Tiago de Oliveira. "Bivariate extremes; extensions". In: *Bull. Inst. Internat. Statist.* 46.2 (1975). With discussion, 241–252, 253–254 (1976).

[168] J. Tiago de Oliveira. "Extremal distributions". In: *Rev. Fac. Sci. Lisboa, Ser. A 7* (1958), pp. 215–227.

[169] J. Tiago de Oliveira. "Extremal distributions". In: *Rev. Fac. Sci. Lisboa*. Vol. 7. Ser.A.

[170] J. Tiago de Oliveira. "La représentation des distributions extrémales bivariées". In: *Bull. Inst. Internat. Statist.* 39.livraison 2 (1962), pp. 477–480.

[171] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288. ISSN: 00359246. URL: http://www.jstor.org/stable/2346178.

[172] R. J. Tibshirani. "The lasso problem and uniqueness". In: *Electron. J. Statist.* 7 (2013), pp. 1456–1490. DOI: 10.1214/13-EJS815. URL: https://doi.org/10.1214/13-EJS815.

[173] J. A. Tropp. "Column subset selection, matrix factorization, and eigenvalue optimization". In: *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics. 2009, pp. 978–986.

[174] J. A. Tropp. "Norms of random submatrices and sparse approximation". In: *Comptes Rendus Mathematique* 346.23 (2008), pp. 1271–1274.

[175] A. M. TURING. "I.—COMPUTING MACHINERY AND INTELLIGENCE". In: *Mind* LIX.236 (1950), pp. 433–460. DOI: 10.1093/mind/LIX.236.433. eprint: /oup/backfile/content_public/journal/mind/lix/236/10.1093_mind_lix.236.433/1/433.pdf. URL: http://dx.doi.org/10.1093/mind/LIX.236.433.

[176] A. W. van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998, pp. xvi+443. ISBN: 0-521-78450-6.

[177] S. A. Van De Geer, P. Bühlmann, et al. "On the conditions used to prove oracle results for the Lasso". In: *Electronic Journal of Statistics* 3 (2009), pp. 1360–1392.

[178] R. Varga. *Geršgorin and His Circles*. Vol. 36. Jan. 2004.

[179] J. L. Wadsworth and J. A. Tawn. "Efficient inference for spatial extreme value processes associated to log-Gaussian random functions". In: *Biometrika* 101.1 (2014), pp. 1–15. ISSN: 0006-3444. DOI: 10.1093/biomet/ast042. URL: http://dx.doi.org/10.1093/biomet/ast042.

[180] J. L. Wadsworth and J. A. Tawn. "Efficient inference for spatial extreme value processes associated to log-Gaussian random functions". In: *Biometrika* 101.1 (2014), pp. 1–15. DOI: 10.1093/biomet/ast042. eprint: /oup/backfile/content_public/journal/biomet/101/1/10.1093_biomet_ast042/2/ast042.pdf. URL: http://dx.doi.org/10.1093/biomet/ast042.

[181] R. H. Walden. "Analog-to-digital converter survey and analysis". In: *IEEE Journal on Selected Areas in Communications* 17.4 (Apr. 1999), pp. 539–550. ISSN: 0733-8716. DOI: 10.1109/49.761034.

[182] S. Weisberg. *Applied Linear Regression*. Third. Hoboken NJ: Wiley, 2005. URL: http://www.stat.umn.edu/alr.

[183] E. T. Whittaker. "XVIII.—On the Functions which are represented by the Expansions of the Interpolation-Theory". In: *Proceedings of the Royal Society of Edinburgh* 35 (1915), pp. 181–194. DOI: 10.1017/S0370164600017806.

[184] L. Wolf and A. Shashua. "Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach". In: *Journal of Machine Learning Research* 6.Nov (2005), pp. 1855–1887.

[185] T. W. Yee. *Vector Generalized Linear and Additive Models: With an Implementation in R*. 1st. Springer Publishing Company, Incorporated, 2015. ISBN: 9781493928170.

[186] T. W. Yee and Alec G. Stephenson. "Vector generalized linear and additive extreme value models". In: *Extremes* 10.1 (June 2007), pp. 1–19. ISSN: 1572-915X. DOI: 10.1007/s10687-007-0032-4. URL: https://doi.org/10.1007/s10687-007-0032-4.

[187] P. Youssef. "Restricted invertibility and the Banach–Mazur distance to the cube". In: *Mathematika* 60.01 (2014), pp. 201–218.

[188] Z. Zhao and H. Liu. "Spectral feature selection for supervised and unsupervised learning". In: *Proceedings of the 24th international conference on Machine learning*. ACM. 2007, pp. 1151–1157.