Mélanges bayésiens de modèles d'extrêmes multivariés, Application à la prédétermination régionale des crues avec données incomplètes.

Sabourin Anne

N° d'ordre: 137 - 2013 Année 2013

THÈSE DE L'UNIVERSITÉ DE LYON

Délivrée par

L'UNIVERSITÉ CLAUDE BERNARD LYON 1

École doctorale InfoMaths

DIPLÔME DE DOCTORAT

Mention: Mathématiques et applications

(arrêté du 7 août 2006)

Soutenue publiquement le 24 septembre 2013 par

Anne SABOURIN

Mélanges bayésiens de modèles d'extrêmes multivariés, Application à la prédétermination régionale des crues avec données incomplètes.

Sous la direction de :

Mme Anne-Laure FOUGÈRES, M. Philippe NAVEAU.

Jury :

- M. Anthony DAVISON (rapporteur)
- M. Johan SEGERS (rapporteur)
- Mme Anne-Laure FOUGÈRES
- M. Philippe NAVEAU
- Mme Clémentine PRIEUR
- M. Stéphane ROBIN
- M. Éric SAUQUET

Remerciements

Merci à tous les membres du jury d'avoir pu se libérer aujourd'hui et particulièrement aux rapporteurs, Anthony Davison et Johan Segers, pour leur relecture du manuscrit et leurs retours.

Cette thèse a été financée par le corps des Ponts, des Eaux et des Forêts. Je tiens à remercier ceux de l'ENGREF qui m'ont guidée et soutenue dans cette entreprise, je pense en particulier à Jacques Breger, Cyril Kao, Gabriel Lang et Éric Parent.

Je remercie chaleureusement mes directeurs de thèse, Anne-Laure Fougères et Philippe Naveau, pour ces trois dernières années. Vous avez su m'apporter tous les encouragements et conseils dont j'avais besoin et m'avez fait confiance en m'accordant une grande liberté. Vos convergences et divergences même m'ont beaucoup apporté et travailler avec vous fut un vrai plaisir.

Un grand merci à Rick Katz, Dan Cooley, Eric Gilleland et leurs collègues du NCAR, qui m'ont mis le pied à l'étrier pendant mes séjours à Boulder dont je garde un excellent souvenir. Merci à Benjamin Renard pour son crash-course en hydrologie statistique et cette fructueuse et agréable collaboration, à Michel Lang pour l'avoir rendue possible, enfin merci à tous ceux que j'ai eu la chance de rencontrer lors des divers séminaires et conférences, et qui m'ont fait avancer, scientifiquement et humainement. Je pense entre autres à Anne Schindler, Mathieu Ribatet, Clément Dombry, Nicolas Eckert, à l'équipe MIA de l'AgroParisTech, aux participants et organisateurs des journées Asterisk, aux rencontres AppliBugs, au workshop à Ascona, à Rochebrune, aux rencontres Lyon-Grenoble, au récent workshop à Copenhague, aux séminaires des thésards de Télécom et du LSTA, et à la joyeuse équipe de Saint-Flour ...

J'ai passé d'agréables moments au LSCE, grâce à l'équipe ESTIMR et aux occupants du bâtiment 701, permanents ou de passage, pendant les pauses café du mardi. Merci à Jérome et Julien pour leur tentative presque réussie de conversion au touch rugby, à Jean-Yves pour les dégustations arctiques, à Benjamin pour les commentaires sur les Dirichhhlettes, à Pradeebane pour son calme olympien ...Enfin merci Théo d'avoir vaillamment supporté ta co-bureaute dans ses moments de doute et de jubilation. Je garderai aussi de très bons souvenirs de mes séjours à Lyon, grâce à la gentillesse et la bonne humeur indestructible des occupants du bureau 111a, d'Alexis et d'Élodie.

Un grand merci à Jean Bérard pour le coup de pouce sur les MCMC.

Je voudrais remercier tous ceux que j'ai plus ou moins ennuyés en leur parlant de stats, d'inondations et de mesure invariante, c'est à dire tous mes amis, pour m'avoir de temps en temps distraite de ces palpitantes questions. Merci Fanny d'avoir été là, comme toujours. Merci Pierre de t'être moqué de mes malheurs et merci Cha d'avoir traversé la Manche quelquefois. Merci Claire-A pour les pauses déjeuner et tes Aventures Extraordinaires. Merci Stacha pour les bols d'air des MichaMichauv's et descentes au flambeau ventées, merci à Mathilde et Greffons associés pour les apéros du Vendredi et le reste et merci aux (ex) grimpeurs et assimilés, parisiens et crestois.

Enfin je remercie toute ma famille, pour, faisons court, tout. Une mention spéciale à Papa Maman pour leur patience et leur foi indestructible en chérie-chérie, et à Marc et Laurence pour tous les rôti-purée-décompressez.

Alex, grand combattant de la pomme de pin, merci. Tu es pour beaucoup dans tout ceci.

Résumé

La théorie statistique univariée des valeurs extrêmes se généralise au cas multivarié mais l'absence d'un cadre paramétrique naturel complique l'inférence de la loi jointe des extrêmes. Les marges d'erreur associées aux estimateurs non paramétriques de la structure de dépendance sont difficilement accessibles à partir de la dimension trois. Cependant, quantifier l'incertitude est d'autant plus important pour les applications que le problème de la rareté des données extrêmes est récurrent, en particulier en hydrologie. L'objet de cette thèse est de développer des modèles de dépendance entre extrêmes, dans un cadre bayésien permettant de représenter l'incertitude. Le chapitre 2 explore les propriétés des modèles obtenus en combinant des modèles paramétriques existants, par mélange bayésien (Bayesian Model Averaging). Un modèle semi-paramétrique de mélange de Dirichlet est étudié au chapitre suivant : une nouvelle paramétrisation est introduite afin de s'affranchir d'une contrainte de moments caractéristique de la structure de dépendance et de faciliter l'échantillonnage de la loi a posteriori. Le chapitre 4 est motivé par une application hydrologique : il s'agit d'estimer la structure de dépendance spatiale des crues extrêmes dans la région cévenole des Gardons en utilisant des données historiques enregistrées en quatre points. Les données anciennes augmentent la taille de l'échantillon mais beaucoup de ces données sont censurées. Une méthode d'augmentation de données est introduite, dans le cadre du mélange de Dirichlet, palliant l'absence d'expression explicite de la vraisemblance censurée. Les perspectives sont discutées au chapitre 5.

Mots clés : Extrêmes multivariés, dépassements de seuil, Bayesian Model Averaging, modèles de mélange, mélanges de Dirichlet, échantillonnage MCMC, augmentation de données, prédétermination des crues.

Bayesian model mergings for multivariate extremes

Application to regional predetermination of floods with incomplete data

Abstract

Uni-variate extreme value theory extends to the multivariate case but the absence of a natural parametric framework for the joint distribution of extremes complexifies inferential matters. Available non parametric estimators of the dependence structure do not come with tractable uncertainty intervals for problems of dimension greater than three. However, uncertainty estimation is all the more important for applied purposes that data scarcity is a recurrent issue, particularly in the field of hydrology.

The purpose of this thesis is to develop modeling tools for the dependence structure between extremes, in a Bayesian framework that allows uncertainty assessment. Chapter 2 explores the properties of the model obtained by combining existing ones, in a Bayesian Model Averaging framework. A semi-parametric Dirichlet mixture model is studied next: a new parametrization is introduced, in order to relax a moments constraint which characterizes the dependence structure. The re-parametrization significantly improves convergence and mixing properties of the reversible-jump algorithm used to sample the posterior. The last chapter is motivated by an hydrological application, which consists in estimating the dependence structure of floods recorded at four neighboring stations, in the 'Gardons' region, southern France, using historical data. The latter increase the sample size but most of them are censored. The lack of explicit expression for the likelihood in the Dirichlet mixture model is handled by using a data augmentation framework.

Keywords: Multivariate extremes, threshold excesses, Bayesian Model Averaging, mixture models, Dirichlet mixtures, MCMC sampling, data augmentation, predetermination of floods.

Productions de la thèse

- **Publications**
  - Sabourin, A., Naveau, P. et Fougères, A.-L., Bayesian model averaging for multivariate extremes, *Extremes*, 2013, pp 1-26.
  - Sabourin, A. et Naveau, P., Bayesian Dirichlet mixture model for multivariate extremes : a re-parametrization, *à paraître dans Computational statistics and Data Analysis.*
  - Sabourin, A., Semi-parametric modelling of excesses above high multivariate thresholds with censored data, *soumis pour publication.*
  - Sabourin, A. et Renard, B., Combining regional estimation and historical floods : a multivariate semi-parametric peaks-over-threshold model with censored data, *en préparation.*
- **Packages** utilisables en `R`
  - `BMAmevt` : Implémentation du Bayesian Model Averaging pour les extrêmes multivariés.

    (Disponible sur `CRAN`[1])
  - `DiriXtremes` : Modèle de mélange de Dirichlet pour la structure de dépendance d'extrêmes multivariés, à nombre variable de composants, avec un algorithme à sauts réversibles échantillonnant la loi a posteriori.

    (À soumettre prochainement, disponible sur demande)
  - `DiriCens` : Adaptation des méthodes d'inférence dans le modèle de Dirichlet au cas des données censurées.

    (À soumettre prochainement, disponible sur demande)

[1] page web : `http://cran.r-project.org/`

Table des matières

Chapitre 1

Introduction

1.1 Contexte et objectifs de la thèse

Estimer la probabilité d'occurrence d'un événement exceptionnel jamais observé, définir un événement correspondant à une faible probabilité d'occurrence, ou caractériser le comportement du maximum d'un échantillon de grande taille sont des problèmes intervenant dans de nombreux domaines d'application, parmi lesquels on peut citer l'assurance, les télécommunications, ou la gestion des risques industriels et environnementaux. La théorie des valeurs extrêmes permet d'estimer de telles grandeurs sous certaines conditions de régularité de la loi des observations.

L'inférence univariée est un problème désormais bien balisé, s'appuyant sur le cadre paramétrique fourni par l'asymptotique. Le résultat fondamental de Fisher et Tippett (1928) détermine les distributions limites possibles de maxima d'échantillons de grande taille : de telles lois sont entièrement caractérisées par un paramètre de forme, auquel s'ajoutent un paramètre d'échelle et un de localisation. Le livre de Gumbel (1958) est le premier ouvrage de référence destiné aux applications dans l'ingénierie, avec une attention particulière accordée au risque de crue (« The flood problem »). On cite généralement la thèse de doctorat de de Haan (1970) comme un des textes fondateurs de la théorie des valeurs extrêmes. Un cadre probabiliste et statistique cohérent s'est par ailleurs développé. Entre autres, Leadbetter *et al.* (1983) étendent les résultats aux situations de dépendance temporelle ; les travaux de Resnick (1987) mettent l'accent sur la caractérisation des extrêmes sous l'angle de la variation régulière et des processus ponctuels. Citons également, sans exhaustivité, Embrechts *et al.* (1997), Beirlant *et al.* (2004), de Haan et Ferreira (2006), ou l'ouvrage introductif de Coles (2001) parmi les contributeurs à l'unification de la théorie, à sa diffusion et au développement de méthodes d'inférence adaptées à différentes applications.

La théorie s'étend au cas multidimensionnel, et à sa généralisation infinidimensionnelle que constituent les processus max-stables et les processus de Pareto généralisés. Ceci permet d'aborder des questions liées aux probabilités

de dépassements simultanés de seuils élevés par une grandeur d'intérêt mesurée en plusieurs points, ou par plusieurs quantités en un point. Par exemple, la « qualité de l'air » est une notion qui dépend de la concentration atmosphérique de différents polluants. C'est bien souvent la combinaison de fortes concentrations au cours d'une même journée qui pose le plus de problèmes de santé publique. De même, pour la prévention du risque hydrologique, il peut être nécessaire de connaître la probabilité d'occurrence de crues affectant simultanémant plusieurs cours d'eau d'une même région. En termes statistiques, ces différents problèmes reviennent à calculer la probabilité d'atteinte d'une « *failure region* » (zone sensible) par une grandeur multivariée ou une fonction aléatoire. Une question intimement liée est celle du comportement joint des maxima d'une telle grandeur, calculés point par point sur une longue période. Un exemple typique d'application concerne la construction d'ouvrages de protection aux Pays-Bas : il s'agit de construire une digue qui, avec une probabilité proche de 1, ne soit dépassée en aucun point par le niveau maximum de la mer sur une très longue période.

Dans la plupart des applications, en particulier en hydrologie, la nécessité de ne garder que les plus grandes observations pour l'inférence restreint le jeu de données et peut induire une forte incertitude. Dans un cadre inférentiel bayésien, des intervalles de crédibilité sont définis même pour des échantillons de petite taille et accessibles par échantillonnage. L'inférence bayésienne n'est pas une nouveauté dans le domaine des extrêmes. Coles et Powell (1996) passent en revue une série de problèmes univariés pouvant être abordés sous un angle bayésien. Pour les applications à l'hydrologie, toujours en univarié, Parent et Bernier (2003) traitent le problème de la modélisation bayésienne des dépassements de seuils élevés (POT, *Peaks-over-threshold*). en utilisant des données historiques. Dans le contexte de l'analyse régionale des crues, Ribatet *et al.* (2007) s'intéressent à l'incorporation de l'information disponible aux sites voisins dans la loi a priori des paramètres du site d'intérêt. Pour la modélisation spatiale des précipitations extrêmes, Cooley *et al.* (2007) adoptent une approche bayésienne hiérarchique des paramètres des lois univariées décrivant les extrêmes en chaque point et Blanchet et Davison (2011) proposent un modèle bayésien paramétrique, max-stable, pour décrire la dépendance des maxima annuels de hauteurs de neige en Suisse.

Dans tous ces exemples, l'inférence a lieu au sein d'un modèle paramétrique, c'est-à-dire qu'on estime un nombre fini de paramètres caractérisant entièrement la distribution d'intérêt. Cependant, contrairement à son analogue univarié, la théorie multivariée et infini-dimensionnelle ne fournit aucun cadre paramétrique pour la structure de dépendance des extrêmes. Se restreindre à un modèle paramétrique pour son inférence revient donc à faire une hypothèse forte et, bien souvent, à ignorer l'incertitude attachée au choix du modèle (l'erreur de modélisation). Il paraît donc raisonnable de se placer dans un cadre non paramétrique ou, à défaut, dans un modèle paramétrique très flexible, autrement dit, dans un modèle avec peu d'hypothèses structu-

relles. Il existe des estimateurs non paramétriques classiques de la structure de dépendance (Einmahl *et al.* (2001); Einmahl et Segers (2009) ou plus récemment de Carvalho *et al.* (2013)), mais leur variance asymptotique n'a pas d'expression explicite. Avoir une représentation de l'incertitude attachée aux estimateurs est pourtant primordial dans un contexte où, en l'absence d'hypothèses fortes de départ et avec peu de données à disposition, cette dernière est susceptible d'être importante. Dans un cadre bayésien non paramétrique, il n'existe pour l'instant que le modèle de Guillotte *et al.* (2011), uniquement applicable aux problèmes bivariés, et le modèle de mélange de Dirichlet, proposé par Boldi et Davison (2007), sur lequel nous reviendrons en détail.

L'objectif méthodologique de cette thèse est de contribuer au développement de modèles aussi flexibles que possible pour la structure de dépendance d'extrêmes multivariés, dans un cadre bayésien permettant de représenter l'incertitude a posteriori. D'un point de vue pratique, ces modèles doivent pouvoir répondre à des questions posées par l'hydrologie, en particulier dans le domaine de la prédétermination des crues. L'approche privilégiée consiste à modéliser les dépassements simultanés de seuils élevés, par opposition à la modélisation des maxima par blocs. Ainsi, les grandeurs manipulées correspondent à des événements pouvant être effectivement observés, alors que les maxima ponctuels sont le résultat, sauf exception, d'événements ayant eu lieu à des dates différentes.

Plusieurs modèles paramétriques multivariés existent déjà. Puisqu'aucun n'est parfait, ils peuvent de manière générale, ajustés sur un même jeu de données, produire des estimations différentes et la question du choix du meilleur modèle émerge naturellement. Plutôt que d'éliminer tous les modèles sauf un, on peut aussi chercher à prendre en compte l'ensemble des estimations. Le cadre bayésien est adapté pour cela, car il permet d'affecter des poids *a priori* à chaque modèle et de les actualiser au vu des observations, de manière à obtenir des poids *a posteriori*. Ces derniers seront affectés aux estimations de chaque modèle pour produire un estimateur moyenné sur l'union des modèles. Cette approche, dite de *Bayesian model averaging* (BMA), que l'on peut traduire par « combinaison bayésienne de modèles » permet ainsi d'élargir le cadre paramétrique en considérant l'union des différents modèles à disposition. L'adaptation de cette méthode au contexte des extrêmes multivariés est l'objet du chapitre 2, et de l'article « Bayesian model averaging for multivariate extremes », paru dans la revue *Extremes* (Sabourin *et al.*, 2013). Le BMA comporte un certain nombre de limitations. D'une part, il ne permet pas d'augmenter infiniment la dimension du modèle, sauf si les modèles à moyenner forment une famille infinie de modèles emboîtés, ce qui n'est pas le cas ici : on reste dans un cadre paramétrique. D'autre part, dès que le jeu de données atteint une taille raisonnable (une centaine de points pour des problèmes de dimension inférieure à cinq), un phénomène général de concentration de la loi a posteriori affecte presque toute la masse à un seul

des modèles en concurrence, de sorte que le résultat est le même que si l'on avait pratiqué une sélection de modèle. On n'obtient donc pas de nouvelle famille paramétrique, même si l'on réduit l'incertitude attachée au choix de modèle.

La possibilité d'une famille infinie de modèles emboîtés n'est pas hypothétique : un modèle de mélange dans lequel on autorise un nombre arbitraire de composants en est un exemple, abondamment présent dans les méthodes d'estimation par noyaux. Dans le cadre POT, les excès multivariés sont caractérisés par une *mesure angulaire*, c'est-à-dire par la distribution des « directions » dans lesquelles les excès ont lieu, autrement dit, par la loi des « proportions relatives » des différentes composantes du vecteur aléatoire d'intérêt, conditionnellement au dépassement d'un seuil défini globalement. La structure de dépendance peut ainsi être caractérisée par une mesure de probabilité définie sur le simplexe. C'est dans ce contexte qu'un mélange de Dirichlet a été proposé par Boldi et Davison (2007). Le modèle est dense dans la classe des mesures angulaires admissibles et, d'un point de vue inférentiel, le nombre de composants du mélange peut être laissé libre d'évoluer au cours de l'exécution de l'algorithme d'échantillonnage (« reversible-jump algorithm », Green (1995)). Pour être valide, la mesure angulaire doit satisfaire des contraintes de moments, ce qui se traduit, dans le mélange de Dirichlet, par une contrainte barycentrique sur les centres des noyaux de densité et constitue une difficulté majeure pour l'inférence. Le problème apparaît particulièrement lors de l'étape de simulation de paramètres du mélange au cours de l'échantillonnage de la loi a posteriori. Au delà de la dimension deux, la chaîne de Markov converge très lentement vers sa distribution stationnaire. Une alternative étudiée par Boldi (2004) dans sa thèse de doctorat consiste à revenir à un cadre fréquentiste et à utiliser un algorithme EM (Expectancy-maximization) pour maximiser la vraisemblance, mais là encore, la variance asymptotique n'est pas accessible. Le chapitre 3 de cette thèse est principalement constitué de l'article « Bayesian Dirichlet mixture model for multivariate extremes : a re-parametrization », accepté pour publication dans la revue *Computational Statistics and Data Analysis* en 2013 (Sabourin et Naveau, 2013). Une nouvelle paramétrisation du modèle de mélange y est proposée, dans laquelle la contrainte de moments est automatiquement satisfaite. Ceci facilite la construction d'un algorithme à sauts réversibles dont les propriétés sont étudiées d'un point de vue théorique et sur des simulations. L'inférence est ainsi rendue possible dans un cadre bayésien, en dimension modérée (de l'ordre de cinq).

Le dernier volet de cette thèse est motivé par un problème d'hydrologie. La « prédétermination des crues » fait référence à l'analyse statistique de données de débit de rivières permettant d'établir la fréquence d'événements rares ou d'estimer des niveaux de débit correspondant à une fréquence donnée. Le problème général de la rareté des données est particulièrement sensible dans ce domaine, en raison de la faible durée des périodes d'observa-

tions disponibles, c'est-à-dire bien souvent une cinquantaine d'années, alors que la grandeur d'intérêt est typiquement la crue centennale, voire millénale. Les arguments basés sur l'appartenance à un domaine d'attraction ne faisant pas l'unanimité pour de telles échelles temporelles, il est naturel de chercher à prendre en compte plus de données. Le cas d'étude est le suivant : Neppel *et al.* (2010) ont reconstruit des débits en quatre points de la région des Gardons cévenols, dans le sud de la France, à partir d'archives recensant les crues historiques. Dans cette thèse, on travaillera avec les « débits de pointe » journaliers, c'est à dire avec les maxima journaliers de débit enregistrés en chaque station. Ces données sont en grande partie censurées : on sait par exemple que le débit de pointe a dépassé un certain seuil ou n'a pas dépassé un autre. Pour fixer les idées, les quatre séries temporelles de débits, avec les intervalles de censure, sont représentés en figure 1.1. La première crue a lieu à Alès en 1609, aucune donnée n'étant disponible sur les autres stations. Ensuite, aucun événement extrême n'est enregistré jusqu'en 1741, de sorte que les débits marginaux à Alès sont considérés comme inférieurs à celui de 1604, et les débits sur les autre stations sont manquants. La « période historique » s'étend jusqu'en 1892, date à laquelle des relevés systématiques de débits de pointe journaliers sont effectués, avec des périodes d'interruption et des domaines d'invalidité de la courbe de tarage permettant de convertir les hauteurs en débits. À partir des années 1970 à Saint-Jean et Mialet, et seulement à partir de 2008 pour Alès et Anduze, des mesures automatiques horaires de débit sont disponibles, dont le maximum journalier peut être extrait.

La connaissance de la loi jointe des extrêmes a d'autant plus d'intérêt en hydrologie, qu'elle permet d'avoir recours à des méthodes de régionalisation de certains paramètres marginaux, toujours afin d'augmenter le nombre de données disponibles par paramètre à estimer. Une analyse univariée à partir des maxima annuels a été conduite par Neppel *et al.* (2010), la modélisation de la structure de dépendance est l'objet du chapitre 4 de cette thèse. À cette occasion, le modèle de mélange de Dirichlet est adapté à une situation de censures multivariées. La figure 1.2 montre les projections bivariées des données mesurées aux quatre stations. À première vue, les extrêmes semblent fortement dépendants : de nombreux points sont présents sur la diagonale, loin de l'origine, ce qui indique que les forts débits ont tendance à apparaître simultanément. De plus, la situation paraît asymétrique : la quantité d'information et l'intensité de la dépendance varient d'une paire à l'autre. Ainsi, un modèle flexible, n'imposant pas de symétrie particulière dans la structure de dépendance paraît adapté et le transfert d'information rendu possible par la modélisation d'une telle structure est susceptible de modifier les estimations sur les stations les moins bien renseignées.

Le problème de la vraisemblance censurée, sans expression explicite, est traité par l'introduction d'un schéma d'augmentation de données, présenté et étudié en section 4.1 dans l'article « Semi-parametric modelling of ex-

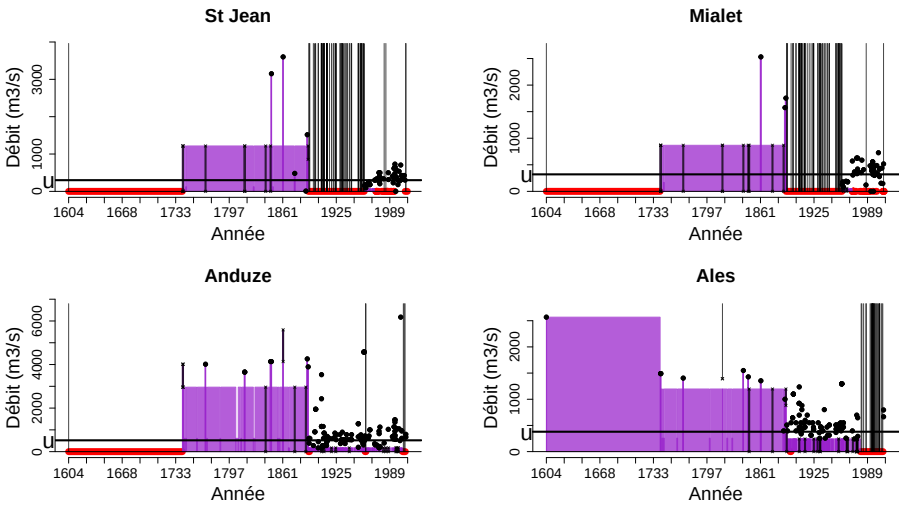

FIGURE 1.1 – Extraction des dépassements de seuil (ligne horizontale) de débit de pointe journaliers aux quatre stations de mesure considérées. Les segments et les aires en violet représentent les données censurées à gauche et/ou à droite, avant extraction des clusters. Les points rouges sont les jours « manquants ». Les segments gris représentent les jours censurés (ou manquants) appartenant à un cluster (correspondant à un dépassement de seuil sur une des quatre stations)

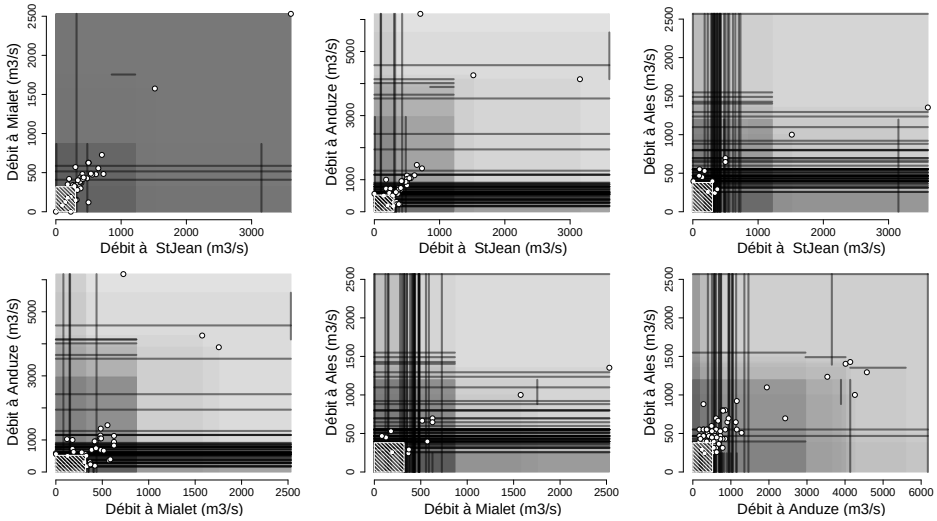

FIGURE 1.2 – Représentation bivariées des dépassements de seuil (rectangle hachuré à l'origine) enregistrés aux six paires de stations, entre 1604 et 2010. Les points correspondent aux données « exactes » ; les segments (*resp.* rectangles) correspondent aux données censurées selon une (*resp.* deux) coordonnée(s). La superposition des données censurées est représentée par des niveaux de gris de plus en plus foncés.

cesses above high multivariate thresholds with censored data », soumis pour publication. Les données des Gardons sont analysées à part : la section 4.2 est constituée d'une version préliminaire de l'article « Combining regional estimation and historical floods : a multivariate semi-parametric peaks-over-threshold model with censored data », co-écrit avec Benjamin Renard (IRSTEA, Institut National de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture, Lyon, France). Ce dernier article sera soumis très prochainement à une revue d'hydrologie.

Les sections suivantes de ce chapitre exposent les éléments de la théorie des valeurs extrêmes et de la statistique bayésienne qui sont utilisés dans la suite. Les chapitres suivants sont organisés autour des quatre articles écrits au cours de cette thèse, les conclusions et perspectives ouvertes sont discutées dans le dernier chapitre.

1.2 Théorie des valeurs extrêmes

Dans cette section, on introduit le cadre général des modèles d'extrêmes multivariés. On s'intéresse en particulier aux dépassements de seuils élevés et c'est d'abord en ces termes que sont exposés les éléments de la théorie univariée indispensables pour la suite. On introduit également le modèle de Poisson utilisé dans le dernier chapitre pour prendre en compte des données censurées avec des seuils de censures variables.

Notations

Dans la suite, les vecteurs sont écrits en caractères gras, la notation $x_{1:n}$ signifie « le vecteur $\boldsymbol{x} = (x_1, \ldots, x_n)$ ». Sauf mention contraire, $\mathbf{R}$ désigne l'ensemble des nombres réels, $\mathbf{N}$ celui des entiers naturels, $(\cdot)_+$ est la partie positive d'une quantité réelle. Si $\boldsymbol{E}$ est un espace topologique, $\mathcal{E}$ désigne l'ensemble des boréliens de $\boldsymbol{E}$. Dans un espace vectoriel, on note $t.A = \{x : \frac{1}{t}.x \in A\}$. En pratique, l'espace des observations sera le plus souvent une partie de $\mathbf{R}^d$, l'espace vectoriel réel canonique de dimension d, ou de sa compactification d'Alexandroff $\bar{\mathbf{R}}^d = [\infty, \infty]^d$. La mesure de Lebesgue sur $\mathbf{R}^d$ ou sur tout sous espace approprié, selon le contexte, est notée ℓ. L'opérateur « maximum » est noté $\bigvee$, les opérateurs arithmétiques et $\bigvee$ appliqués à des vecteurs sont définis composante par composante. Les relations de comparaison entre vecteurs s'entendent comme le produit des comparaisons marginales, c'est-à-dire $\boldsymbol{X} \leq \boldsymbol{x} \Leftrightarrow (X_1 \leq x_1) \cap \cdots \cap (X_d \leq x_d)$. L'abréviation *f.d.r.* et le terme « loi » signifient « fonction de répartition ». Si F est une fonction de répartition, alors $\bar{F} = 1 - F$ est la fonction de survie associée. La convergence faible des mesures de probabilité, notée $\xrightarrow{d}$, correspond à l'intégration contre les fonctions continues bornées $(\mathcal{C}_b)$: une suite de mesures de probabilité (P_n) sur $\boldsymbol{E}$ converge faiblement vers P_0 si et seulement si, pour

toute fonction $f \in \mathcal{C}_b(\boldsymbol{E})$, $\int_{\boldsymbol{E}} f \, \mathrm{d}P_n \to \int_{\boldsymbol{E}} f \, \mathrm{d}P_0$.

1.2.1 Limite des excès et des maxima : cas univarié

De nombreux ouvrages exposent en détail les bases de la théorie des valeurs extrêmes, par exemple de Haan et Ferreira (2006), Beirlant *et al.* (2004), Resnick (1987), Leadbetter *et al.* (1983). Cette dernière référence, en particulier, présente et prouve de manière relativement concise les résultats qui suivent en partant de l'étude des maxima. Les formes possibles des lois limites des excès sont obtenues dans Balkema et De Haan (1974), et Pickands (1975).

Soit Y une variable aléatoire réelle, de loi F. On note $\omega(F) \in \mathbf{R} \cup \{\infty\}$ la borne supérieure du support de F. On cherche à caractériser la loi des excès au-dessus de seuils u élevés, c'est-à-dire la loi de la variable conditionnelle $(Y-u)|Y>u$. Dans le cas $\omega(F) < \infty$, *i.e.* lorsque Y est bornée supérieurement, « pour les seuils élevés » signifie « pour les seuils proches de $\omega(F)$ ».

La loi conditionnelle des excès au-dessus du seuil $u < \omega(F)$ est

$$F_u(x) := P(Y - u \leq x \,|Y > u) = \frac{F(x+u) - F(u)}{1 - F(u)}, \quad x \geq 0\,.$$

La fonction de survie conditionnelle $\bar{F}_u = 1 - F_u$ s'exprime plus simplement :

$$\bar{F}_u(x) := P(Y - u > x \,|Y > u) = \frac{\bar{F}(u+x)}{\bar{F}(u)}, \quad x \geq 0\,.$$

Pour obtenir une limite en loi, on s'autorise une re-normalisation affine des excès, en considérant la variable

$$\tilde{Y}_{u,\sigma_u,\mu_u} := \frac{(Y-u) - \mu_u}{\sigma_u} \;\mid (Y > u)\;,$$

de fonction de survie

$$P(\tilde{Y}_{u,\sigma_u,\mu_u} > x) = \bar{F}_u(\sigma_u x + \mu_u)\,, \quad x > -\mu_u/\sigma_u\,,\; \sigma_u > 0,\; \mu_u \in \mathbf{R}\,.$$

L'hypothèse indispensable permettant d'étudier les grandes valeurs de Y dans le cadre de la théorie des valeurs extrêmes est l'existence d'une variable aléatoire Z, de loi Q non dégénérée, et de constantes $\sigma_u > 0, \mu_u \in \mathbf{R}$, définies pour $u < \omega(F)$, telles que

$$\tilde{Y}_{u,\sigma_u,\mu_u} \xrightarrow[u \to \omega(F)]{d} Z\,, \tag{1.1}$$

ce qui signifie

$$\forall x \in \mathscr{C}(Q), \quad \bar{F}_u(\sigma_u x + \mu_u) \xrightarrow[u \to \omega(F)]{} \bar{Q}(x)\,, \tag{1.2}$$

où $\mathscr{C}(Q)$ désigne l'ensemble des points de continuité de Q.

On dit alors que F appartient au domaine d'attraction des excès de Q ($F \in \mathrm{DAE}(Q)$), et (1.1) ou (1.2) est appelée dans la suite *condition du domaine d'attraction des excès* (DAE). Cette hypothèse, qui peut paraître arbitraire au premier abord, et en réalité assez « raisonnable » : elle est vérifiée par la plupart des lois de probabilité continues usuelles. C'est le cas, par exemple, de la loi normale, des lois Gamma, des lois « à queues lourde » comme la loi de Cauchy, ou de la loi Bêta.

En étudiant la limite de $\bar{F}_u(\sigma_{\mu_u} x + \mu_{\mu_u})$ et en utilisant le théorème de convergence des types de Khintchine, énoncé ci-dessous, (1.1) implique que Q est *stable par seuillage*, c'est-à-dire qu'il existe, pour tout u tel que $0 < Q(u) < 1$, des constantes α_u et β_u, telles que, pour tout point de continuité x de Q,

$$Q_u(\alpha_u(x) + \beta_u) = Q(x)\,. \tag{1.3}$$

Réciproquement, si Q est stable par seuillage, elle est dans son propre domaine d'attraction car il y a égalité dans (1.2) en choisissant $\sigma_u = \alpha_u$, $\mu_u = \beta_u$.

On dit que deux *f.d.r.* F et F^* sont de même type s'il existe des constantes $a > 0$, b, telles que $F(\,\cdot\,) = F^*(a(\,\cdot\,) + b)$. La propriété de stabilité par seuillage de Q signifie donc que Q et Q_u sont de même type. Le théorème suivant établit l'égalité en type entre les les limites de *f.d.r.* obtenues par différentes normalisations affines d'une même *f.d.r.* initiale.

Théorème 1 (Convergence des types, Khintchine)**.** *Soit F_n et F_n^* deux suites de* f.d.r. *sur* $\mathbf{R}$*, et G une* f.d.r. *non dégénérée. Soient $a_n > 0$, b_n une suite de constantes telles que*

$$\forall x \in \mathscr{C}(G), \quad F_n(a_n\, x + b_n) \xrightarrow[n\to\infty]{} G(x)\,.$$

et soient $\alpha_n > 0$, β_n d'autres suites de constantes de normalisation et G^* une f.d.r. *elle aussi non dégénérée.*

Alors,

$$\forall x \in \mathscr{C}(G^*), \quad F_n(\alpha_n\, x + \beta_n) \xrightarrow[n\to\infty]{} G^*(x)\,.$$

si et seulement s'il existe des constantes $A > 0$, B, telles que

$$\frac{\alpha_n}{a_n} \to A\,, \quad \frac{\beta_n - b_n}{a_n} \to B\,,$$

et l'on a alors

$$\forall x \in \mathbf{R}, \quad G^*(x) = G(A\,x + B)\,.$$

Pour une preuve, voir *e.g.* Resnick (1987), chapitre 0, ou Leadbetter *et al.* (1983), chapitre 1.

Le résultat principal utilisé dans la modélisation POT est le suivant.

Théorème 2 (Balkema, de Haan, 1974). *Si Y satisfait une condition du domaine d'attraction* (1.1)*, alors la limite $\bar{Q}$ dans* (1.2) *est du type*

$$\bar{Q}_\xi(x) = (1+\xi x)_+^{\frac{-1}{\xi}} \ , \quad x \geq 0 \, ,$$

où le membre de droite s'interprète, pour $\xi = 0$, comme sa limite lorsque ξ tend vers 0*, c'est-à-dire $\bar{Q}_0(x) = e^{-x}$.*

Ce résultat s'obtient en résolvant une équation fonctionnelle de type Hamel sur $\varphi = \log(\bar{Q})$, elle même obtenue en utilisant le Théorème 1. L'uniformité de la convergence sur les intervalles bornés inférieurement est établie dans Pickands (1975).

Autrement dit, la famille des limites possibles dans (1.2), appelées « lois de Pareto généralisées » (GPD, *Generalized Pareto Distribution*) est l'ensemble des

$$\{Q_{\xi,\sigma,\mu} : \xi \in \mathbf{R}, \sigma > 0, \mu \in \mathbf{R}\}$$

définies par

$$\bar{Q}_{\xi,\sigma,\mu}(x) = \begin{cases} \left(1+\xi\,\frac{x-\mu}{\sigma}\right)_+^{-\frac{1}{\xi}} & (\frac{x-\mu}{\sigma} > 0) \\ 1 & (\frac{x-\mu}{\sigma} < 0) \end{cases} \tag{1.4}$$

Le paramètre ξ est appelé *paramètre de forme*, μ et σ ont les paramètres de localisation et d'échelle. Le paramètre de forme détermine la forme du support de Q_ξ : pour $\xi \geq 0$, $\mathrm{supp}(Q_\xi) = [0, \infty]$; le support est borné pour $\xi < 0$: $\mathrm{supp}(Q_\xi) = [0, -1/\xi]$.

L'énoncé 2 est le pendant du résultat fondamental de Fisher et Tippett (1928) sur les limites possibles des maxima renormalisés de suites *i.i.d.* de variables aléatoires. Historiquement, c'est en effet par une condition de convergence sur les maxima que l'on caractérise l'appartenance de F à un domaine d'attraction. Pour la généralisation au cas multivarié, il sera plus simple d'énoncer une condition de convergence des maxima ponctuels plutôt que des excès. Soient $Y_n, n \geq 1$ des variables indépendantes, de loi F. La condition sur les maxima est qu'il existe une variable aléatoire Z, de loi G non dégénérée, et de constantes $a_n > 0, b_n \in \mathbf{R}$, telles que

$$\frac{\bigvee_{t=1}^n Y_t - b_n}{a_n} \xrightarrow[n\to\infty]{d} Z \ , \tag{1.5}$$

ce qui signifie

$$\forall y \in \mathscr{C}(G), F^n(a_n y + b_n) \xrightarrow[n\to\infty]{} G(y) \, . \tag{1.6}$$

De manière similaire à la condition d'attraction sur les excès, on dit alors que F appartient au domaine d'attraction de G ($F \in \mathrm{DA}(G)$) et (1.5) ou (1.6) est la *condition du domaine d'attraction* (DA). On appelle *loi de valeurs extrêmes*, ou loi GEV (*Generalized Extreme Value distribution*) toute

fonction de répartition limite G dans (1.6). En étudiant cette fois-ci la limite de $F^n(a_{nk} \cdot +b_{nk})$ pour tout $k \in \mathbf{N}$, et toujours par un argument de convergence des types, une loi de valeurs extrêmes G est nécessairement *max-stable*, c'est-à-dire qu'il existe, pour tout $k \in \mathbf{N}$, des constantes α_k et β_k, telles que

$$\forall y \in \mathbf{R}\,,\ G^k(\alpha_k\, y + \beta_k) = G(y)\,. \tag{1.7}$$

En utilisant la monotonie de G, on montre facilement que (1.7) a encore lieu en prenant k dans $\mathbf{R}^+$. Le résultat fondamental de la théorie univariée, qui s'obtient par exemple dans de Haan (1976) et Leadbetter *et al.* (1983) en résolvant une équation fonctionnelle sur la réciproque du logarithme itéré de G, $(-\log(-\log(G)))^{\leftarrow}$, est le suivant.

Théorème 3 (Fisher and Tippett (1928), Gnedenko (1943)).
La limite G dans (1.6) est de type

$$G_\xi(y) = \exp\big[-(1+\xi\, y)_+^{-\frac{1}{\xi}}\big]\,, \quad \xi \in \mathbf{R}\}\,, \tag{1.8}$$

où le membre de droite s'interprète, pour $\xi = 0$, comme sa limite lorsque ξ tend vers 0, c'est-à-dire $G_0(y) = \exp\big[-e^{-y}\big]$.

En comparant (1.4) et (1.8), on voit que les lois limites des maxima sont les exponentielles des fonctions de survie limites des excès : si Q est une GPD, alors $\exp(-\bar{Q})$ est une GEV.

Les lois GEV se classifient, comme les GPD, par le signe de ξ. Pour $\xi = 0$, le support de G est $\bar{\mathbf{R}}$ et un représentant canonique est la *loi de Gumbel*, G_0, obtenue avec $\sigma = 1$, $\mu = 0$. Pour $\xi > 0$, le support est borné inférieurement, et G a le type d'une *loi de Fréchet*, $\Phi_\alpha(y) = e^{-y^{-\alpha}}$ $(y > 0, \alpha > 0)$. À l'inverse, pour $\xi < 0$, le support est borné supérieurement et G est du même type qu'une *loi de Weibull*, $\Psi_\alpha(y) = e^{-(-y)^\alpha}$ $(y < 0, \alpha > 0)$. Notons que les lois limites des excès sont dans le domaine d'attraction des lois de valeurs extrêmes de même paramètre de forme. La figure 1.3 montre les densités et les fonctions de répartitions correspondant aux trois types de lois GEV. Elles sont comparées avec la densité et la *f.d.r.* de la loi normale, qui est dans le domaine d'attraction de Gumbel. Le terme « loi à queues lourdes », réservé aux distributions dans le domaine de Fréchet, fait référence à la décroissance lente de la queue de distribution.

La figure 1.4 montre trois exemples de trajectoires de suites de variables aléatoires indépendantes, respectivement distribuées selon une loi de Fréchet, de Weibull et de Gumbel. Comme prévu, les variables de Fréchet sont bornées inférieurement et prennent de très grande valeurs en de rares occasions ; les variables de Weibull sont bornées supérieurement et approchent de très près la borne supérieure, enfin les variables de Gumbel ne sont pas bornées mais sont sujettes à de moins fortes anomalies que celles de Fréchet.

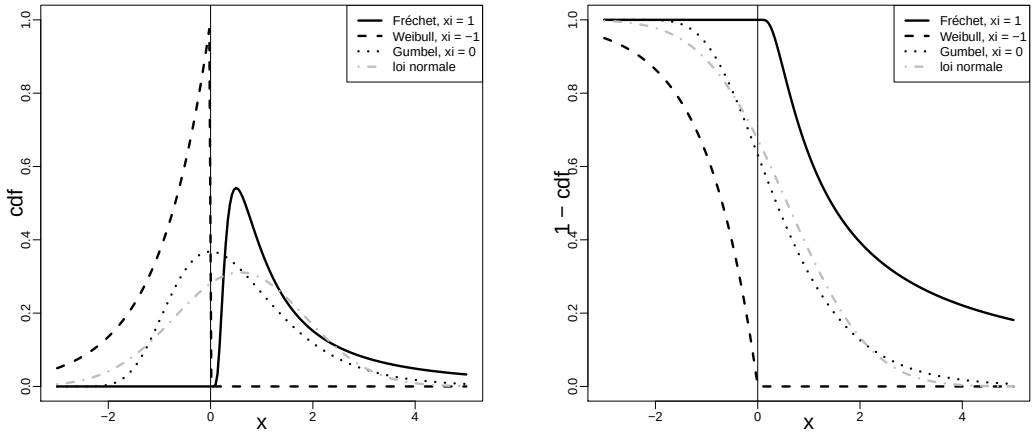

FIGURE 1.3 – Distributions de valeurs extrêmes ; densité et fonction de survie pour les trois types, comparées à la loi normale.

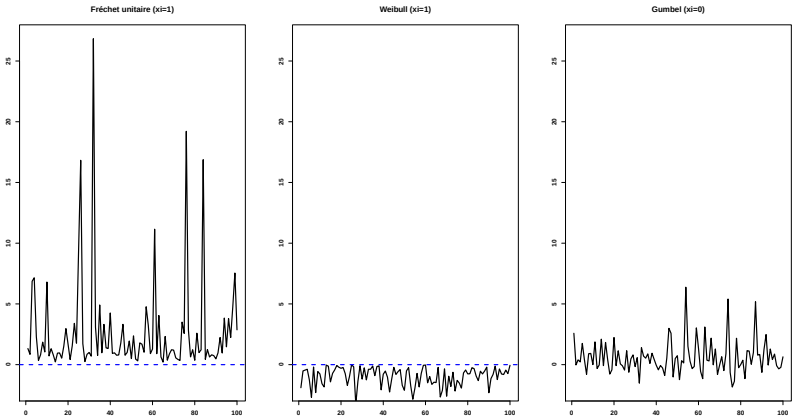

FIGURE 1.4 – Simulation de variables indépendantes, distribuées selon une loi de Fréchet (gauche), Weibull (milieu) et Gumbel (droite)

Équivalence des conditions de domaine d'attraction Le résultat suivant établit le lien entre les deux approches et indique que l'on peut prendre $\mu_u \equiv 0$ dans (1.2).

Théorème 4 (Balkema et De Haan (1974))**.** *Se valent :*

1. $F \in \mathrm{DAE}(Q_\xi)$
2. $F \in \mathrm{DA}(G_\xi)$ *avec des constantes de normalisations* $a_n > 0$, b_n, *et l'on a*
$$G_\xi = \exp\left[-\bar{Q}_\xi\right]$$
3. *Il existe une fonction strictement positive* σ, *telle que*
$$\frac{\bar{F}(u + x\,\sigma(u))}{\bar{F}(u)} \xrightarrow[u \to \omega(F)]{} \bar{Q}_\xi(x))\,. \tag{1.9}$$

De plus, en posant $a(u) = a_{\lfloor u \rfloor}$*, on peut choisir* $\sigma(u) = a\big(\frac{1}{1-F(u)}\big)$.

Ceci s'obtient par exemple dans le Théorème 1.1.6, de Haan et Ferreira (2006), en étudiant la convergence d'une version renormalisée de $U = (\frac{1}{1-F})^{\leftarrow}$, où $(f)^{\leftarrow}$ est la fonction inverse généralisée, continue à gauche, de la fonction croissante f.

Autrement dit, (1.5) équivaut à ce que la loi de Y, conditionnellement à un excès au-dessus du seuil u élevé, soit approximativement une GPD. C'est la raison d'être du modèle suivant, largement employé dans la littérature (par exemple, par Davison et Smith, 1990; Coles et Tawn, 1991) et utilisé dans toute la suite pour l'estimation des lois marginales au sein des modèles d'extrêmes multivariés :

$$F^{\xi,\sigma}(x) = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \leq x} & (x \leq u) \\ 1 - \zeta_u \left(1 + \xi \frac{x-u}{\sigma}\right)^{-1/\xi} & (x > u)\,, \end{cases}$$

où $(Y_i)_{1 \leq i \leq n}$ est un échantillon identiquement distribué, et ζ_u est la probabilité d'excès du seuil u. En règle générale, on choisit u comme un quantile élevé, mais suffisamment peu pour que la variance de l'estimateur empirique soit faible. On néglige alors l'erreur d'estimation. Ce modèle est souvent qualifié de « semi-paramétrique » car la partie « sub-seuil » du modèle est non paramétrique (la *f.d.r.* empirique) alors que la « sup-seuil » est paramétrique, de type GPD.

Domaine d'attraction de Fréchet Ce domaine joue un rôle particulier dans la suite, car la loi de Fréchet est choisie comme *f.d.r.* standard pour les marginales univariées (voir section 1.2.2). Une notion centrale pour l'étude de ce domaine d'attraction est la suivante.

Définition 1 (Variation régulière)**.**
Une fonction f *est dite à variation régulière s'il existe une fonction strictement positive* $h : \mathbf{R} \to \mathbf{R}$*, telle que*

$$\frac{f(tx)}{f(t)} \xrightarrow[t\to\infty]{} h(x) \quad x > 0\,.$$

h *vérifie alors* $h(x) = x^{\rho}$*, pour un certain* $\rho \in \mathbf{R}$.

Si $\rho \neq 0$*,* f *est dite à variation régulière d'indice* ρ *et l'on note* $f \in \mathrm{RV}(\rho)$ *cette propriété.*

Dans le cas $\rho = 0$*,* f *est dite à* variation lente.

La forme particulière de h s'obtient du fait qu'elle vérifie l'équation de Hamel $h(xy) = h(x)\,h(y)$ (voir, *e.g.* Resnick, 1987, Proposition 0.4)

Il y a équivalence entre variation régulière et condition du domaine d'attraction de Fréchet. Plus précisément, on a

Proposition 1 (Gnedenko (1943)).

$$F \in \mathrm{DA}(\Phi_\alpha) \Leftrightarrow \bar{F} \in \mathrm{RV}(-\alpha) \tag{1.10}$$

Dans ce cas, on peut prendre comme constantes dans (1.6)

$$a_n = \left(\frac{1}{\bar{F}}\right)^{\leftarrow}(n)\,; \quad b_n = 0\,.$$

On obtient ce résultat par usage répété d'un théorème de Karamata, selon lequel une fonction U est à variation régulière d'indice $\rho \geq -1$, si et seulement si

$$\frac{x\,U(x)}{\int_0^x U(t)\,\mathrm{d}t} \xrightarrow[x\to\infty]{} \rho + 1\,,$$

et en travaillant sur la fonction $U = (\frac{1}{1-F})^{\leftarrow}$.

1.2.2 Cas multivarié

Comme dans le cas univarié, l'hypothèse minimale est l'existence de vecteurs de normalisation $(\boldsymbol{a}_n) = (a_{1,n}, \ldots, a_{d,n}), (\boldsymbol{b}_n) = (b_{1,n}, \ldots, b_{d,n})$, avec $a_{j,n} > 0$, et d'une *f.d.r.* G non dégénérée, tels que

$$F^n(\boldsymbol{a}_n \boldsymbol{y} + \boldsymbol{b}_n) \xrightarrow[n\to\infty]{} G(\boldsymbol{y})\,. \tag{1.11}$$

Ici encore, les limites possibles sont les lois max-stables : (1.11) a lieu pour au moins une *f.d.r.* F si et seulement s'il existe, pour tout $t > 0$, des vecteurs $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$, avec $\alpha_{j,t} > 0$ $(1 \leq j \leq d)$, telles que

$$\forall \boldsymbol{y} \in \mathbf{R}^d, \quad G^t(\boldsymbol{\alpha}_t \boldsymbol{y} + \boldsymbol{\beta}_t) = G(\boldsymbol{y})\,.$$

La caractérisation de telles lois nécessite quelques définitions et l'introduction de la topologie vague des mesures. Soit $\boldsymbol{E}$ un espace métrique complet, séparable (CMS), localement compact, typiquement $\boldsymbol{E} = \mathbf{R}^d$, et soit $\mathcal{C}_c^+$ l'ensemble des fonctions continues, positives, à support compact sur $\boldsymbol{E}$. Une mesure borélienne sur $\boldsymbol{E}$, finie sur les compacts, est dite *de Radon*. On note $M_+(\boldsymbol{E})$, ou tout simplement M_+, l'ensemble des mesures de Radon positives sur $\boldsymbol{E}$.

Topologie vague La topologie privilégiée pour étudier la convergence dans M_+ est la *topologie vague*, dont les ouverts sont obtenus par intégration contre les fonctions continues à support compact, c'est-à-dire que les ensembles

$$\left\{\lambda \in M_+ \,:\, s < \int_{\boldsymbol{E}} f\,\mathrm{d}\lambda \,< t\right\}\,,$$

pour $s, t \in \mathbf{R}$ et $f \in \mathcal{C}_c^+$, forment une pré-base d'ouverts pour cette topologie. Ainsi, par définition, une suite $(\lambda_n)_{n\geq 1} \in M_+(\boldsymbol{E})$ converge vaguement vers λ_0

si et seulement si, pour toute $f \in \mathcal{C}_c^+$, $\lambda_n(f) := \int_{\boldsymbol{E}} f \, \mathrm{d}\lambda_n \longrightarrow \lambda_0(f)$. On note alors $\lambda_n \xrightarrow{v} \lambda_0$. L'intérêt technique de la topologie vague est que, muni de cette topologie, M_+ est polonais, *i.e.* séparable et métrisable pour une métrique qui le rend complet. Il y a donc, par le théorème de Prohorov, équivalence entre tension d'une suite de mesure et existence d'un point d'accumulation.

Comme pour la convergence faible, on a un théorème portemanteau (*e.g.* Resnick, 1987, prop. 3.12)

Proposition 2 (Portemanteau, convergence vague)**.** *Il y a équivalence entre*

1. $\lambda_n \xrightarrow{v} \lambda$ *sur* $\boldsymbol{E}$ *;*
2. *Pour tout* $B \in \mathcal{E}$ *tel que* $\lambda(\partial B) = 0$, $\lambda_n(B) \xrightarrow[n\to\infty]{} \lambda(B)$ *;*
3. *Pour tout compact* K, $\limsup \lambda_n(K) \leq \lambda(K)$ *et pour tout ouvert* U *d'adhérence compacte,* $\liminf \lambda_n(U) \geq \lambda(U)$.

Lois infiniment divisibles Les lois max-stables G sont en particulier *max-infiniment divisibles*, c'est-à-dire que, pour tout $t > 0$, G^t est une fonction de répartition. Les lois max-infiniment divisibles s'écrivent (Resnick, 1987, proposition 5.8)

$$G(\boldsymbol{y}) = \begin{cases} e^{-\lambda[-\boldsymbol{\infty},\boldsymbol{y}]^c} . & \text{si } \boldsymbol{y} \geq \boldsymbol{l} \,, \\ 0 & \text{sinon} \,, \end{cases} \tag{1.12}$$

où $\boldsymbol{l} \in [-\boldsymbol{\infty}, \boldsymbol{\infty}[^d$, peut être choisi comme la borne inférieure du support de G, et où λ est une *mesure exponentielle*, *i.e.*

- λ est une mesure positive de Radon sur $\boldsymbol{E} = [\boldsymbol{l}, \boldsymbol{\infty}] \setminus \{\boldsymbol{l}\}$,
- λ définit bien une *f.d.r.* sur $\mathbf{R}^d$ à travers (1.12), c'est-à-dire : λ ne charge pas les hyperplans passant à l'infini ($\lambda(\boldsymbol{E} \setminus [-\boldsymbol{\infty}, \boldsymbol{\infty})) = 0$), et dans le cas où $\boldsymbol{l} \not> -\boldsymbol{\infty}$, on doit avoir, pour $\boldsymbol{y} \geq \boldsymbol{l}$ et $y_j = -\infty$, $\lambda[-\boldsymbol{\infty}, \boldsymbol{y}]^c = \infty$.

On obtient (1.12) en considérant la suite de mesures $\lambda_n := nG^{\frac{1}{n}}$, et en vérifiant, par un argument de compacité, sa convergence vague sur $\boldsymbol{E}$ vers une limite λ vérifiant nécessairement

$$\lambda([-\boldsymbol{\infty}, \boldsymbol{y}]^c) = \lim_{n\to\infty} n(1 - G^{\frac{1}{n}}(\boldsymbol{y})) = -\log G(y) \,.$$

Représentation des lois max-stables La structure de la loi jointe limite G dans (1.11) apparaît plus clairement si l'on transforme les lois marginales en lois de valeurs extrêmes standard, par exemple des lois de Fréchet unitaires, en posant $X_j = \frac{-1}{\log(F_j(Y_j))}$, de sorte que $\mathrm{P}(X_j \leq x) = e^{-1/x}$, $x > 0$. Soit F_* la *f.d.r.* des variables standardisées, $F_*(\boldsymbol{x}) = \mathrm{P}(\boldsymbol{X} \leq \boldsymbol{x}) = F\left(F_1^{\leftarrow}(e^{-1/x_1}), \ldots, F_d^{\leftarrow}(e^{-1/x_d})\right)$.

D'après Resnick (1987), Proposition 5.10, la condition (1.11) équivaut à ce que

1. Les F_j soient dans le domaine d'attraction de lois max-stables univariées,

2. F_* soit dans le domaine d'attraction d'une loi max-stable multivariée G_*, dont les marges sont Fréchet unitaires (Φ_1).

Par un argument de convergence des types et puisqu'on peut prendre $a_{j,t} = t$, $b_{j,t} = 0$ dans les conditions de DA des lois marginales $(F_*)_j$, on a

$$G_*^t(t.\boldsymbol{x}) = G_*(\boldsymbol{x}).$$

La mesure exponentielle λ_* de G_* est définie sur $\boldsymbol{E} = [\mathbf{0}, \boldsymbol{\infty}]^d \setminus \{\mathbf{0}\}$ (*i.e.* on peut prendre $\boldsymbol{l} = \mathbf{0}$ dans (1.12), car le support des $(G_*)_j$ est $[0, \infty]$), et l'égalité précédente implique

$$t\,\lambda_*(t.A) = \lambda_*(A)\,, \quad A \in \mathcal{E}\,. \tag{1.13}$$

Pour tirer parti de l'homogénéité de λ_*, il est préférable d'utiliser une système de coordonnées polaires. Soit $\|\cdot\|$ une norme sur $\boldsymbol{E}$ et $\boldsymbol{S}_d$ le quadrant positif de la sphère unité. Soit $\mathfrak{S}$ la mesure sur $\boldsymbol{S}_d$ définie par

$$\mathfrak{S}(A) = \lambda_*(\{t.A, t \geq 1\}).$$

La mesure $\mathfrak{S}$ est finie par compacité dans $\boldsymbol{E}$ de $\{t.A, t \geq 1\}$. Soit T la transformation en coordonnées polaires

$$\begin{aligned} T\ :\quad & \boldsymbol{E} \to ((0,\infty] \times \boldsymbol{S}_d) \\ & \boldsymbol{x} \mapsto (r, \boldsymbol{w}) = (\|x\|, \|x\|^{-1}\,.\,\boldsymbol{x})\,. \end{aligned}$$

La propriété d'homogénéité (1.13) s'exprime par le fait que $\lambda_* \circ T^{-1}$ est une mesure produit :

$$\mathrm{d}(\lambda_* \circ T^{-1}) = \frac{1}{r^2}\,\mathrm{d}r \times \mathrm{d}\mathfrak{S}\,, \tag{1.14}$$

Ainsi, $\mathfrak{S}$ contient toute l'information sur la structure de dépendance de G. On l'appelle *mesure spectrale* ou *mesure angulaire*. Le choix initial de lois marginales Fréchet unitaires impose $\int_{\boldsymbol{S}_d} x_j\,\mathrm{d}\mathfrak{S} = 1\,(1 \leq j \leq d)$.

Choisissons la norme L^1 : $\|\boldsymbol{x}\|_1 = \sum_{j=1}^d x_j$. On a alors

$$\int_{\boldsymbol{S}_d} \mathrm{d}\mathfrak{S} = \int_{\boldsymbol{S}_d} (\sum_j x_j)\,\mathrm{d}\mathfrak{S} = d\,.$$

En posant $H = \frac{1}{d}\mathfrak{S}$, H est une mesure de probabilité sur $\boldsymbol{S}_d$, qu'on appelle *mesure de probabilité angulaire* ou, par abus de langage, *mesure angulaire*.

Par intégration de (1.14) sur des pavés cartésiens, la relation entre H et λ_* est

$$\lambda_*[\mathbf{0}, \boldsymbol{x}]^c = d \int_{\boldsymbol{S}_d} \bigvee_{j=1}^{d} \frac{w_j}{x_j} \, \mathrm{d}H(\boldsymbol{w}) \,.$$

On obtient alors la représentation dite de Pickands :

Théorème 5. *Une* f.d.r. *non dégénérée G_* dont les lois marginales $(G_*)_j$ sont Fréchet unitaires, est max-stable si et seulement si elle admet la représentation*

$$G_*(\boldsymbol{x}) = \exp\left[-d \int_{\boldsymbol{S}_d} \bigvee_{j=1}^{d} \frac{w_j}{x_j} \, \mathrm{d}H(\boldsymbol{w})\right] , \tag{1.15}$$

où la mesure angulaire H est une mesure de probabilité sur le simplexe $\boldsymbol{S}_d$ vérifiant

$$\int_{\boldsymbol{S}_d} w_j \, \mathrm{d}H(\boldsymbol{w}) = \frac{1}{d} \,, \quad 1 \leq j \leq d \,. \tag{1.16}$$

La condition de moments (1.16) découle du choix de lois marginales Fréchet unitaires.

Variation régulière Comme dans le cas du DA univarié de Fréchet, la condition sur F_* est liée à la variation régulière de $1 - F_*$. En effet, cette notion s'étend au cas multivarié :

Définition 2 (Variation régulière multivariée)**.** *Une fonction $f : \mathbf{R}^d \to \mathbf{R}$ est dite à variation régulière sur $\mathbf{R}^d$ s'il existe une fonction $h : \mathbf{R}^d \to \mathbf{R}$, telle que*

$$\frac{f(t.\boldsymbol{x})}{f(t.\mathbf{1})} \xrightarrow[n\to\infty]{} h(\boldsymbol{x}) \,.$$

Il existe alors $\rho \in \mathbf{R}$ tel que h est homogène d'indice ρ,

$$h(s.\boldsymbol{x}) = s^\rho \, h(\boldsymbol{x}) \,.$$

L'homogénéité de la limite h est établie par exemple dans Resnick (1987), section 5.4.2., de manière similaire au cas univarié.

L'intérêt de travailler avec des lois marginales standardisées apparaît dans la proposition suivante, qui découle de ce qui précède et des propositions 5.15 et 5.17 de Resnick (1987).

Proposition 3. *Soit F_* une* f.d.r. *dont les marges sont dans* $\mathrm{DA}(\Phi_1)$ *et soit $\boldsymbol{X} \sim F_*$. En coordonnées polaires, soient $R = \|\boldsymbol{X}\|_1$, $\boldsymbol{W} = \frac{1}{R}\boldsymbol{X}$. Il y équivalence entre*

1. *$F_* \in \mathrm{DA}(G_*)$;*
2. *La fonction de survie $1 - F_*$ est à variation régulière ;*

3. $n\mathrm{P}(n^{-1}.\boldsymbol{X} \in \cdot\,) \xrightarrow{v} \lambda_*(\,\cdot\,)$, *où* λ_* *est la mesure exponentielle de* G_* *;*

4. *Si* H *est la mesure de probabilité angulaire associée à la mesure exponentielle* λ_**, on a*

$$\begin{cases} \mathrm{P}(\boldsymbol{W} \in A \,|R > r\,) \xrightarrow[r\to\infty]{} H(A)\,, \quad A \subset \boldsymbol{S}_d, \ \textit{mesurable}, \\ \mathrm{P}(R > r) \underset{r\to\infty}{\sim} \frac{d}{r}\,. \end{cases} \tag{1.17}$$

Dans ce cas, les marges de G_* sont Φ_1, l'indice de variation régulière de $1 - F_*$ est 1 et l'on a

$$\frac{1 - F_*(t.\boldsymbol{x})}{1 - F_*(t.\mathbf{1})} \xrightarrow[t\to\infty]{} \frac{\lambda_*[\mathbf{0}, \boldsymbol{x}]^c}{\lambda_*[\mathbf{0}, \mathbf{1}]^c}\,.$$

D'après la proposition 3, si une loi multivarié est dans le domaine d'attraction d'une loi max-stable, alors, après standardisation des lois marginales en lois de Fréchet unitaires, la distributions des excès au-dessus de seuils élevés est entièrement déterminées par la mesure angulaire H. La figure 1.5 illustre ce résultat pour des extrêmes bivariés, simulés selon une loi produit où la partie radiale est une loi de Pareto standard et la partie angulaire est une mesure angulaire valide, c'est à dire vérifiant la contrainte de moments (1.16). Deux cas sont représentés. Le panneau de droite décrit une situation de faible dépendance asymptotique, où H attribue la majorité de sa masse aux régions à proximité des sommets du simplexe. Dans ce cas, les événements extrêmes dans une direction sont le plus souvent modérés dans l'autre direction. Le panneau de gauche, à l'inverse, montre un cas de de forte dépendance asymptotique, où H se concentre au milieu du simplexe. Cette fois-ci, les valeurs extrêmes dans une direction sont souvent associés à une valeur extrême dans l'autre direction, autrement dit, les extrêmes ont lieu simultanément. Lorsque H est dégénérée en masses de Dirac sur les sommets du simplexe, les extrêmes sont dits *asymptotiquement indépendants*. Tous les autres cas rentrent dans le cadre de la *dépendance asymptotique*. Dans cette thèse, on s'attache à modéliser cette dernière situation. Dans le cas extrême où H est concentrée en masse de Dirac au milieu du simplexe, les extrêmes sont dits *totalement dépendants*.

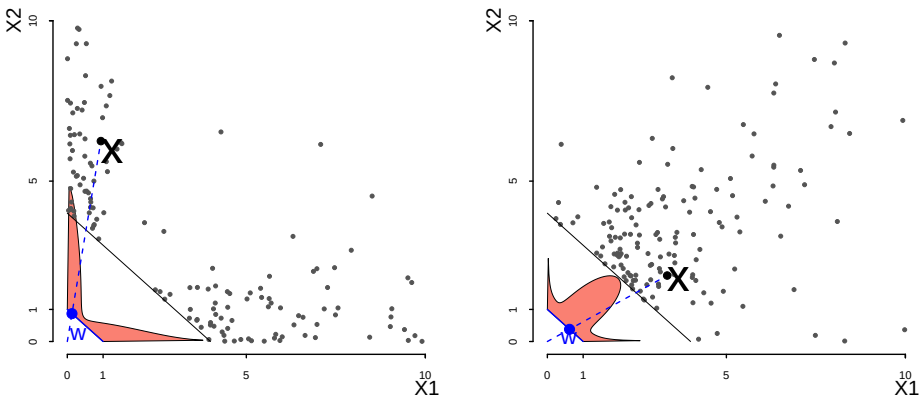

FIGURE 1.5 – Exemples de mesures angulaires bivariées pour les excès au-dessus de seuils radiaux élevés.
Points gris, données simulées. Région rouge pâle, densité de la mesure angulaire. Le point bleu est l'angle $\boldsymbol{W}$ correspondant à l'observation $\boldsymbol{X}$ (point noir).

Mélanges de Dirichlet De manière générale, il n'existe aucune contrainte sur la loi H, à part la condition de moments (1.16) découlant d'un choix de standardisation. Autrement dit, aucune famille paramétrique n'est imposée par l'asymptotique. Cependant, un modèle particulier de mesures angulaires est privilégié aux chapitres 3 et 4 : celui des mélanges de Dirichlet. Une loi de Dirichlet est une mesure de probabilité sur le simplexe $\boldsymbol{S}_d$, de dimension $d-1$, qui généralise la loi Bêta (obtenue pour $d = 2$) aux dimensions supérieures. Elle est caractérisée par un paramètre de localisation $\boldsymbol{\mu} = \mu_{1:d} \in \mathring{\boldsymbol{S}}_d$ et un paramètre de concentration $\nu > 0$. Elle est absolument continue par rapport à la mesure de Lebesgue sur $\{(w_1, \ldots, w_{d-1}) : w_j \geq 0, \sum_{1:d-1} w_j \leq 1\}$, de densité

$$\mathrm{diri}(\boldsymbol{w} \mid \boldsymbol{\mu}, \nu) = \frac{\Gamma(\nu)}{\prod_{i=1}^d \Gamma(\nu\mu_i)} \prod_{i=1}^d w_i^{\nu\mu_i - 1} \qquad \boldsymbol{w} \in \mathring{\boldsymbol{S}}_d\,.$$

Le simplexe $\boldsymbol{S}_d$ étant une représentation de l'ensemble des lois de probabilité sur un ensemble discret de cardinal d, la loi de Dirichlet est bien connue en statistique comme loi sur l'ensemble des lois de probabilité discrètes, de dimension donnée. Elle admet une généralisation infini-dimensionnelle, le processus de Dirichlet, discuté en section 5.2. Les lois et processus de Dirichlet sont largement employés en statistique bayésienne, en tant que lois a priori sur l'espace des paramètres. Notre situation est différente : ici, les lois de Dirichlet sont employées comme lois des observations, plus précisément de leur composante angulaire.

Si $\boldsymbol{W} \sim \mathrm{diri}(\,\cdot\,|\boldsymbol{\mu}, \nu)$, alors $\mathbb{E}(\boldsymbol{W}) = \boldsymbol{\mu}$, ce qui impose, pour obtenir une mesure angulaire valide, $\boldsymbol{\mu} = (1/d, \ldots, 1/d)$. Seul le paramètre ν est libre,

et les deux principales formes possibles obtenues avec $\nu < d$ et $\nu > d$ sont représentées sur la figure 1.6, avec $d = 3$. Le simplexe est alors le triangle équilatéral joignant les exrémités du trièdre formé par les vecteurs de la base canonique sur $\mathbf{R}^3$. Dans le premier cas, la mesure angulaire est amodale et la masse est concentrée vers les bords du simplexe, ce qui correspond à une dépendance asymptotique modérée. Dans le second cas, la mesure est modale et la dépendance asymptotique est forte.

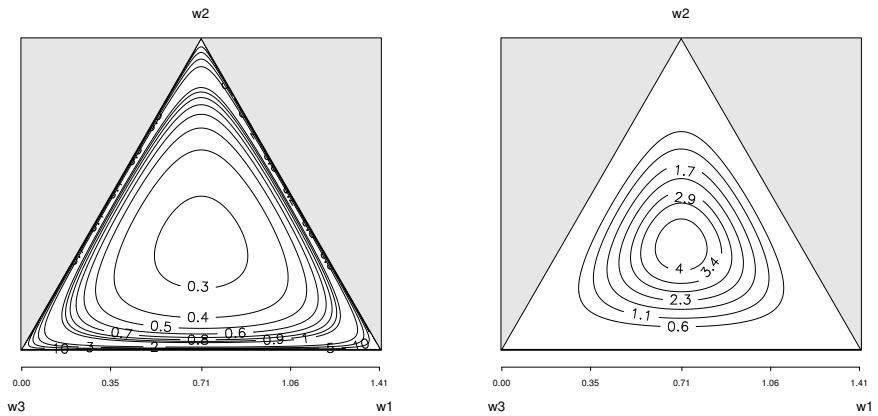

FIGURE 1.6 – Lignes de niveau de densités de Dirichet sur $\boldsymbol{S}_d$ avec $d = 3$, centrées ($\boldsymbol{\mu} = (1/3, 1/3, 1/3)$, pour des paramètre de concentration $\nu = 1$ (gauche) et $\nu = 10$ (droite).

Ce modèle à un paramètre est très peu flexible. En revanche, considérons un mélange de lois de Dirichlet (DM, *Dirichlet mixture*), à k composants, de paramètres $(\boldsymbol{\mu}_{\cdot,1}, \ldots, \boldsymbol{\mu}_{\cdot,k}) := \boldsymbol{\mu}$, $(\nu_1, \ldots, \nu_k) := \boldsymbol{\nu}$, avec des poids $(p_1, \ldots, p_k) := \boldsymbol{p}$, satisfaisant $\sum_{m=1}^{k} p_m = 1$, $p_m > 0$, c'est à dire une densité sur $\boldsymbol{S}_d$ s'écrivant comme une combinaison convexe

$$h_{\boldsymbol{p},\boldsymbol{\mu},\boldsymbol{\nu}}(w) = \sum_{m=1}^{k} p_m \operatorname{diri}(\boldsymbol{w}|\boldsymbol{\mu}_{\cdot,m}, \nu_m)\,, \quad \boldsymbol{w} \in \overset{\circ}{\boldsymbol{S}}_d\,.$$

La contrainte (1.16) est alors équivalente à une contrainte barycentrique s'appliquant aux paramètres de localisation $\boldsymbol{\mu}_{\cdot,1:k}$:

$$(1.16) \Leftrightarrow \sum_{m=1}^{k} p_m\, \boldsymbol{\mu}_{\cdot,m} = (1/d, \ldots, 1/d)\,. \tag{1.18}$$

La figure donne un exemple de densité de mélange satisfaisant (1.18). Le modèle DM est aussi flexible que possible, à condition de laisser varier le nombre k de composants du mélange : plus précisément, Boldi et Davison (2007) ont montré sa densité, au sens faible, dans l'ensemble des mesures angulaires valides.

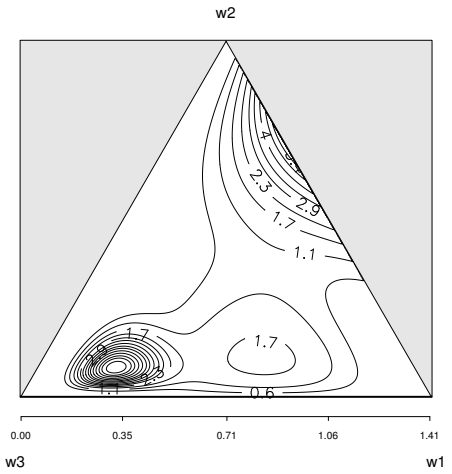

FIGURE 1.7 – Mélange de trois lois de Dirichlet satisfaisant la contrainte de moments (1.18).

Une difficulté majeure rencontrée au stade de l'inférence bayésienne est liée à (1.18) : des paramètres satisfaisant cette contrainte doivent être générés à chaque pas de l'algorithme MCMC. Le principal objet du chapitre 3 est de traiter ce problème en re-paramétrant le modèle, de sorte que la contrainte (1.16) soit automatiquement vérifiée.

Outre sa grande flexibilité, un intérêt du modèle DM tient aux propriétés de marginalisation et de conditionnement des lois de Dirichlet. Ces propriétés sont tout particulièrement exploitées au chapitre 4 pour traiter les données partiellement manquantes ou censurées. Dans le cas d-varié, soit $\boldsymbol{X} = R.\boldsymbol{W}$ un vecteur aléatoire distribué selon la « loi des extrêmes », où $W \sim H$ et R a pour densité r_0/r^2, $r > r_0$, pour un certain $r_0 > 0$. Considérons la loi marginale d'-dimensionnelle de $\tilde{\boldsymbol{X}} = (X_1, \ldots, X_{d'})$, $d' < d$. En termes angulaires, la grandeur d'intérêt est l'angle $\tilde{\boldsymbol{W}}$ correspondant au vecteur marginal $\tilde{\boldsymbol{X}}$. La mesure exponentielle marginale $\tilde{\lambda}$ associée en vertu de la proposition 3, est obtenue par intégration de la densité de λ dans les directions $(d'+1, \ldots, d)$, en coordonnées cartésiennes. Cette dernière densité est liée à la densité h par (Coles et Tawn, 1991, Théorème 1),

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\boldsymbol{x}}(\boldsymbol{x}) = d\,.\,r^{-(d+1)} h(\boldsymbol{w})\,, \quad r = \sum_{j=1}^{d} x_j,\ \boldsymbol{w} = \boldsymbol{x}/r\,.$$

Lorsque h est un mélange de lois de Dirichlet, on a ainsi

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\boldsymbol{x}}(\boldsymbol{x}) = d \sum_{m=1}^{k} \left\{ \frac{p_m \Gamma(\nu_m)}{\prod_{j=1}^{d} \Gamma(\nu_m \mu_{j,m})} \prod_{j=1}^{d} x_j^{\nu_m \mu_{j,m} - 1} \left(\sum_{j=1}^{d} x_j \right)^{-(\nu_m + 1)} \right\}. \quad (1.19)$$

En intégrant (1.19) selon les directions $(d'+1,\dots,d)$, on obtient

$$\begin{aligned}\frac{\partial\tilde{\lambda}(\boldsymbol{x})}{\partial x_1\cdots\partial x_{d'}} &= \int_{(\mathbf{R}^+)^{d-d'}} \frac{\mathrm{d}\lambda}{\mathrm{d}\boldsymbol{x}}(x_1,\dots,x_{d'},x_{d'+1},\dots,x_d)\,\mathrm{d}x_{d'+1}\cdots\mathrm{d}x_d \\ &= d'\sum_{m=1}^{k}\left(\frac{\tilde{p}_m\Gamma(\tilde{\nu}_m)}{\prod_{j\le d'}\Gamma(\tilde{\nu}_m\tilde{\mu}_{j,m})}\prod_{j\le d'}x_j^{\tilde{\nu}_m\tilde{\mu}_{j,m}-1}\left(\sum_{j\le d'}x_j\right)^{-(\tilde{\nu}_m+1)}\right),\end{aligned} \tag{1.20}$$

avec

$$\begin{gathered}\tilde{\nu}_m = \nu_m(\sum_{j\le d'}\mu_{j,m})\ , \quad \tilde{\boldsymbol{\mu}}_{1:d',m} = (\sum_{j\le d'}\mu_{j,m})^{-1}\boldsymbol{\mu}_{1:d',m}\ , \\ \tilde{p}_m = \frac{d}{d'}\sum_{j\le d'}\mu_{j,m}p_m\ .\end{gathered} \tag{1.21}$$

Par identification dans (1.19), la mesure angulaire marginale $\tilde{H}$ correspondant à $\tilde{\lambda}$ est encore un mélange de Dirichlet, de paramètres de mélange $(\tilde{\boldsymbol{p}},\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\nu}})$ donnés par (1.21). Notons que les poids du mélange sont modifiés, de sorte que $\tilde{H}$ vérifie à nouveau la contrainte (1.18).

Ceci contraste avec le cadre classique, dans lequel une loi de Dirichlet est la loi d'une mesure de probabilité aléatoire. La « marginale » d'une loi de mélange de Dirichlet sur les coordonnées $(1,\dots,d')$ est alors définie commec la loi du vecteur $\boldsymbol{W}' = (\sum_1^{d'}W_j)^{-1}\cdot(W_1,\dots,W_{d'})\in\boldsymbol{S}_{d'}$. C'est à nouveau un mélange de Dirichlet, de paramètres $\boldsymbol{\mu}'=\tilde{\boldsymbol{\mu}}$ et $\boldsymbol{\nu}'=\tilde{\boldsymbol{\nu}}$ comme dans (1.21), cependant les poids du mélange sont inchangés : $\boldsymbol{p}'=\boldsymbol{p}\neq\tilde{\boldsymbol{p}}$.

Concernant les lois conditionnelles, toujours dans le cas d-varié, considérons $[X_1|X_2=x_2,\dots,X_d=x_d]$ la loi de la première coordonnée du « vecteur extrémal » $\boldsymbol{X}$ sachant les suivantes. Notons $X_{[1|2:d]}$ la variable conditionnelle. On peut montrer (*c.f* chapitre 4, section 4.1, appendice A), que $X_{1|2:d}$ est distribué selon un mélange de quotients de lois Bêta

$$Y_m \overset{d}{=} \frac{\sigma U_m}{1-U_m} \quad (1\le m\le k)\,,$$

où $U_m \sim \mathrm{beta}(a_m,b_m)$, $a_m=\nu_m\mu_{1,m}$, $b_m=\nu_m(1-\mu_{1,m})+1$ et $\sigma=\sum_{j\ge2}x_j$, avec des poids de mélange

$$p_m^{1|2:d} = p_m\,(1-\mu_{1,m})\frac{\Gamma(\nu_m(1-\mu_{1,m}))}{\prod_{j\ge2}\Gamma(\nu_m\mu_{j,m})}\rho_m\,\sigma^{-\nu_m(1-\mu_{1,m})-1},$$

où $\rho_m=\prod_{j\ge2}x_j^{\nu_m\mu_{j,m}-1}$, de sorte que, si l'on se donne une variable aléatoire $Z\in\{1:k\}$,distribuée selon une loi multinomiale de paramètres $p_m^{1|2:d}$, $1\le m\le k$, on a

$$X_{1|2:d} \overset{d}{=} \sum_{m=1}^{k}\mathbb{1}_{(Z=m)}\frac{\sigma U_m}{1-U_m}\,.$$

Le fait de connaitre la distribution d'une variable sachant toutes les autres permet en particulier de simuler des coordonnées manquantes et d'éviter l'intégration numérique de la vraisemblance en présence de censures. Une autre application serait la simulation conditionnelle d'extrêmes, mais cette possibilité n'est pas explorée dans cette thèse.

Dans le cadre de données censurées, la présence d'intervalles de censure recouvrant le seuil d'« extrémalité » *i.e.* le seuil au-delà duquel le modèle extrémal est censé s'appliquer, empêche d'écrire la vraisemblance dans un modèle de dépassements de seuil (POT) classique. Ce problème peut être contourné (*c.f.* chapitre 4) en utilisant un modèle de Poisson à la place du modèle POT. C'est l'objet du paragraphe suivant.

Processus de Poisson Soit $(\boldsymbol{E}, \mathcal{E})$ un espace CMS localement compact. On munit $M_+(\boldsymbol{E})$ de sa tribu borélienne $\mathcal{M}_+$. Une mesure aléatoire de Radon est une fonction mesurable d'un espace probabilisé $(\Omega, \mathcal{A}, \mathrm{P})$ dans $(M_+, \mathcal{M}_+)$. Un processus ponctuel $\boldsymbol{E}$ est une mesure aléatoire de Radon, à valeurs mesure ponctuelle, c'est-à-dire s'écrivant $\sum_{i=1}^{\infty} \mathbb{1}_{x_i}$, où le nombre de x_i est fini sur tout compact.

Définition 3. *Un processus ponctuel N est appelé* processus de Poisson *sur $\boldsymbol{E}$, d'intensité $\lambda \in M_+(\boldsymbol{E})$, si*

1. *Pour $A \in \mathcal{E}$, $k \in \mathbf{N}$,*

$$\mathrm{P}(N(A) = k) = \begin{cases} \frac{\lambda(A)^k}{k!} e^{-\lambda(A)} & si\ \lambda(A) < \infty\,, \\ 0 & sinon. \end{cases}$$

2. *Pour toute famille $A_1, \ldots, A_r \in \mathcal{E}$ d'ensembles disjoints, les variables aléatoires $N(A_1), \ldots, N(A_r)$ sont mutuellement indépendantes.*

On note alors $N = \mathrm{PP}(\lambda)$.

La convergence faible d'un processus de Poisson est définie comme la convergence faible d'une variable aléatoire à valeurs dans l'espace métrique $M_+(\boldsymbol{E})$. Elle est caractérisée par la convergence des lois fini-dimensionnelles, c'est-à-dire (*c.f.* Daley et Vere-Jones, 2007, théorème 11.1.7), $N_n \xrightarrow{d} N_0$ si et seulement si l'on a la convergence faible dans $\mathbf{R}^p$

$$(N_n(A_1), \ldots N_n(A_p)) \xrightarrow{d} (N_0(A_1), \ldots N_0(A_p))\,,$$

pour tout $p \in \mathbf{N}$ et $A_1, \ldots, A_p \in \mathcal{E}$, tels que $\mathrm{P}(N_0(\partial A_i) > 0) = 0$, $1 \leq i \leq p$. Il est toutefois plus commode d'utiliser une caractérisation par la convergence de la transformée de Laplace. Cette dernière est définie, pour une mesure aléatoire ν, par

$$\mathrm{Lap}_\nu(f) = \mathbb{E}\left(e^{-\int f \,\mathrm{d}\nu}\right)\,, \quad f \geq 0,\ \text{mesurable}.$$

en particulier, pour un processus ponctuel $N = \sum_i \mathbb{1}_{X_i}(\,\cdot\,)$,

$$\mathrm{Lap}_N(f) = \mathbb{E}\left(e^{-\sum_i f(X_i)}\right),.$$

La transformée de Laplace d'une mesure aléatoire détermine entièrement sa loi (Resnick, 1987, prop. 3.5) et l'on a (Resnick, 1987, prop. 3.19) :

Proposition 4. *Une suite $(\nu_n)_{n\geq 1}$ de mesures aléatoires converge faiblement vers une mesure aléatoire ν_0 si et seulement si $\mathrm{Lap}_{(\nu_n)}(f) \to \mathrm{Lap}_{(\nu_0)}(f)$, pour toute fonction f mesurable positive.*

L'intérêt de la variation régulière des fonctions de survie obtenue dans la proposition 3 apparaît à ce stade : elle permet d'obtenir une équivalence entre condition de DA et convergence du processus ponctuel des excès vers un processus de Poisson. Plus précisément, en utilisant la transformée de Laplace, on obtient une équivalence entre convergence vague des mesures empiriques et convergence faible des processus ponctuels empiriques. C'est l'objet de la proposition suivante (Resnick, 1987, proposition 3.21). Ici, ℓ est la mesure de Lebesgue sur $\mathbf{R}$.

Proposition 5. *Soient $(X_{j,n}, j \geq 1)$ des suites de variables aléatoires à valeurs dans $\boldsymbol{E}$,* i.i.d.*pour tout $n \in \mathbf{N}$, et soit N_n le processus empirique sur $\mathbf{R} \times \boldsymbol{E}$,*

$$N_n = \sum_{j\geq 1} \mathbb{1}_{\left(\frac{j}{n},\, X_{j,n}\right)}\,.$$

Soit λ_0 une mesure de Radon sur $\boldsymbol{E}$ et $N_0 = \mathrm{PP}(\ell \times \lambda_0)$ sur $\mathbf{R} \times \boldsymbol{E}$.

Alors,

$$N_n \xrightarrow{d} N_0 \quad \Leftrightarrow \quad n\mathrm{P}(X_{1,n} \in \,\cdot\,) \xrightarrow{v} \lambda_0(\,\cdot\,)\,. \tag{1.22}$$

Les propositions 3 et 5 mènent au résultat final de ce chapitre.

Théorème 6. *On se place dans le cadre de la proposition 3. Chaque assertion qui y est énoncée est équivalente à*

$$\sum_{i=1}^{n} \mathbb{1}_{\left(\frac{t}{n},\, \frac{\boldsymbol{X}_t}{n}\right)}(\,\cdot\,) \xrightarrow[n\to\infty]{d} \mathrm{PP}(\ell \times \lambda_*) \text{ sur } [0,1] \times \boldsymbol{E}\,.$$

Au chapitre 4, l'inférence de la loi jointe des extrêmes sera basée sur l'hypothèse, justifiée par le théorème 6, que le processus ponctuel $\{(\frac{t}{n}, \frac{\boldsymbol{X}_t}{n})\}_{t\leq n}$ formé par les données standardisées $\boldsymbol{X}_t$ (avec des lois marginales Fréchet unitaires) au dessus d'un seuil multivarié, c'est à dire sur une région $(\bar{\mathbf{R}}^+)^d \setminus [0, u_1] \times \cdots [0, u_n]$, est un processus de Poisson d'intensité $\ell \times \lambda_*$. La partie angulaire H de la mesure exponentielle λ_* sera modélisée par un mélange de Dirichlet.

1.3 Statistique bayésienne

La statistique bayésienne a connu un regain d'intérêt ces trente dernières années, avec le développement de moyens de calcul et de méthodes d'échantillonnage rendant applicable la théorie.

Parmi les nombreux ouvrages de référence en la matière, Robert (2007) fournit un grand nombre d'exemples de cas d'étude bayésiens, et discute tout particulièrement les problèmes de théorie de la décision. La monographie de Berger et Wolpert (1988) détaille les connections entre le paradigme bayésien et le principe de vraisemblance. Schervish (1995) est un cours complet de statistique présenté un point de vue principalement bayésien mais exposant en parallèle les approches classiques. Ghosh et Ramamoorthi (2003) est une excellente référence pour qui veut aborder l'inférence non paramétrique. Enfin, Robert et Casella (2004) et Robert et Casella (2010) proposent une introduction aux méthodes numériques bayésiennes. Cette section introduit les notions de base nécessaires à la compréhension des méthodes d'inférence utilisées dans la suite de cette thèse.

1.3.1 Formule de Bayes et loi a posteriori

Dans la suite, $(\boldsymbol{X}, \mathcal{X})$ est l'espace des observations, $\mathcal{P}$ l'ensemble des lois de probabilité sur $\mathcal{X}$. L'espace des paramètres $\boldsymbol{\Theta}$ de dimension finie ou non, et $\mathcal{P}_{\boldsymbol{\Theta}} = \{\mathrm{P}_\theta, \theta \in \boldsymbol{\Theta}\}$ est un sous-ensemble de $\mathcal{P}$, c'est-à-dire un modèle statistique. Les capitales désignent indifféremment les mesures de probabilité ou leur fonction de répartition (s'il y a lieu) et les minuscules sont réservées aux densités.

Rappelons avant tout qu'un noyau de transition sur un produit d'espaces mesurables $(\boldsymbol{\Theta} \times \boldsymbol{X}, \mathcal{T} \otimes \mathcal{X})$, est une famille de probabilités $\{\mathrm{P}_\theta\}_{\theta \in \Theta}$ indexées par Θ, telle que l'application $\theta \mapsto \mathrm{P}_\theta(A)$ soit $\mathcal{T}$-mesurable pour tout $A \in \mathcal{X}$.

Définition 4. *Un modèle statistique bayésien est la donnée de*

1. *Un modèle statistique $\mathcal{P}_{\boldsymbol{\Theta}}$ dominé par une mesure σ-finie ν ;*
2. *Une tribu $\mathcal{T}$ sur $\boldsymbol{\Theta}$, telle que $\{\mathrm{P}_\theta\}_{\theta \in \boldsymbol{\Theta}}$ soit un noyau de transition ;*
3. *Une mesure σ-finie π sur $\mathcal{T}$, appelée* prior, *telle que la mesure* prédictive a priori *$\mu_{\boldsymbol{X}}$, définie par*

$$\mu_{\boldsymbol{X}}(A) = \int_{\boldsymbol{\Theta}} \mathrm{P}_\theta(A) \, \mathrm{d}\pi(\theta), \; A \in \mathcal{X} \,, \tag{1.23}$$

soit σ-finie.

$\mu_{\boldsymbol{X}}$ est aussi appelée « mesure marginale » des observations.

On note alors, pour $x \in \boldsymbol{X}$, $p_\theta(x) = \frac{\mathrm{dP}_\theta}{\mathrm{d}\nu}(x)$ la dérivée de Radon-Nikodym de P_θ par rapport à ν au point x.

Dans le cas où $\pi(\boldsymbol{\Theta}) = 1$, une interprétation probabiliste d'un modèle bayésien consiste à définir le paramètre θ comme une variable aléatoire. Dans ce cas, $\mu_{\boldsymbol{X}}$ est automatiquement σ-finie, c'est même une mesure de probabilité. Si au contraire, $\pi(\boldsymbol{\Theta}) = \infty$, le prior est dit *impropre* mais l'inférence est encore possible tant que $\mu_{\boldsymbol{X}}$ est σ-finie. Choisir un cadre d'inférence bayésien consiste ainsi à munir l'espace des paramètres d'une tribu et d'une mesure, sous des conditions de régularité qui sont vérifiées en dehors des cas pathologiques. En effet, exiger que la prédictive a priori soit σ-finie dans (1.23) revient à imposer au modèle d'être suffisamment discriminant pour que, à domaine d'observation A_i fixé dans un certain recouvrement dénombrable $\bigcup_n A_n$ de $\boldsymbol{X}$, $\mathrm{P}_\theta(A_i)$ décroisse suffisamment vite pour éliminer les valeurs à l'infini de θ.

L'inférence sur θ à partir d'une observation $X = x$ est obtenue par conditionnement du prior.

Loi a posteriori lorsque π est une loi de probabilité Dans ce cas, on peut se donner un espace probabilisé sous-jacent $(\Omega, \mathcal{F}, \mathbb{P})$, et définir $\mu_{\boldsymbol{X}}$ et π comme les mesures images $\mathbb{P} \circ X^{-1}$ et $\mathbb{P} \circ \Theta^{-1}$, où

$$X : \Omega \to \boldsymbol{X} \text{ et } \Theta : \Omega \to \boldsymbol{\Theta}$$

sont des applications mesurables. Par définition du modèle bayésien, on suppose que la probabilité conditionnelle $\mu_{\boldsymbol{X}}(\,\cdot\,|\Theta = \theta)$, a une version régulière $\{P_\theta\}_{\theta \in \boldsymbol{\Theta}}$.

Remarque 1. La fonction d'ensemble aléatoire « probabilité conditionnelle » $A \mapsto \mathbb{P}(X \in A|\Theta) := \mathbb{E}(\mathbb{1}_A(X)|\Theta)$ est toujours définie, et $\mathcal{T}$-mesurable en θ à A fixé, d'après l'existence de l'espérance conditionnelle relativement à la tribu $\sigma(\Theta)$. Il n'est toutefois pas assuré en général (sauf à l'imposer comme au (3) de la définition du modèle bayésien) que la fonction ainsi définie soit bien une mesure aléatoire, car la propriété de σ-additivité peut n'avoir lieu qu'en dehors d'un ensemble négligeable N_θ pouvant dépendre de θ. Pour plus de détails, *c.f.*, par exemple, Kallenberg (1997), chapitre 5.

De même que P_θ est la loi conditionnelle de X sachant $\Theta = \theta$, $\mu_{\boldsymbol{X}}(\,\cdot\,|\Theta = \theta)$, on définit alors la loi a posteriori sur $\boldsymbol{\Theta}$ comme la 'contrepartie' de P_θ, c'est-à-dire la loi conditionnelle de Θ, obtenue en conditionnant la loi π par l'observation x.

Définition 5 (loi a posteriori, prior propre)**.** *Lorsque le prior est propre, on appelle* loi a posteriori *toute version régulière de la probabilité conditionnelle à l'observation, sur l'espace des paramètres, $\pi(\,\cdot\,|X = x)$. Pour $x \in \boldsymbol{X}$, on la note $\pi_x(\,\cdot\,)$ ou $\pi(\,\cdot\,|x)$.*

Cette approche est la plus souvent employée pour introduire la théorie bayésienne ; c'est en particulier celle de Schervish (1995). Elle a l'avantage de donner une interprétation intuitive de la loi a posteriori. Cependant, cette

définition n'a plus de sens dans le cas d'un prior impropre et oblige soit à mener formellement les calculs sans base mathématique rigoureuse, soit à affaiblir les axiomes fondamentaux des probabilités, ce qui force à ré-établir un grand nombre de résultats de base. Pour plus de détails, voir par exemple Schervish (1995), chapitre 1, exemple 1.40, et ses références. Dans cette thèse, seules des lois a priori propres seront utilisées. Cependant, par souci de cohérence avec un cadre plus général, on présente au paragraphe suivant une approche moins restrictive, empruntée à Chang et Pollard (1997), consistant à autoriser dès le départ un prior impropre dans le cadre de la définition 4.

Loi a posteriori lorsque π est impropre La loi a posteriori est alors ce qui se rapproche le plus d'une loi conditionnelle en théorie de la mesure, lorsque la variable de conditionnement n'est pas une variable aléatoire mais un élément d'un espace mesuré plus général, c'est-à-dire une désintégration fibre à fibre de la mesure jointe sur $\boldsymbol{X} \times \boldsymbol{\Theta}$ par rapport à la projection sur $\boldsymbol{X}$, définie comme suit.

Définition 6 (désintégration fibre à fibre (Chang et Pollard, 1997))**.** *Soit $(\mathcal{Y}, \mathscr{Y}, \lambda)$ et $(\boldsymbol{X}, \mathscr{X}, \mu)$ des espace mesurés, λ et μ étant σ-finies.*
Soit $\varphi : \mathcal{Y} \longrightarrow \boldsymbol{X}$, une application mesurable. On appelle désintégration de λ par (φ, μ) toute famille de mesures $\{\lambda_x\}_{x \in \boldsymbol{X}}$ telle que

(i) Pour tout x, λ_x est concentrée sur $\varphi^{-1}(\{x\})$

(ii) Pour toute fonction positive g, λ-mesurable, $x \mapsto \int_{\mathcal{Y}} g \, \mathrm{d}\lambda_x$ est mesurable et

$$\int_{\mathcal{Y}} g \, \mathrm{d}\lambda = \int_{x \in \boldsymbol{X}} \left\{ \int_{\mathcal{Y}} g(y) \, \mathrm{d}\lambda_x(y) \right\} \, \mathrm{d}\mu(x) \, .$$

La mesure μ est appelée mesure de mélange et les λ_x sont les mesures de désintégration.

Proposition 6 (Chang et Pollard (1997), Théorème 2 (iii))**.** *Dans le cas où μ est la mesure image $\lambda \circ \varphi^{-1}$, on parle de « désintégration par φ ». Alors, les mesures de désintégration λ_x sont μ-presque sûrement des mesures de probabilité.*

Cette généralisation de la notion de probabilité conditionnelle permet de définir la loi a posteriori dans le cas général.

Définition 7 (loi a posteriori, prior quelconque)**.** *On se place dans le cadre de la définition 4. Soit $\mathcal{Y}$ l'espace produit $(\boldsymbol{X} \times \boldsymbol{\Theta})$, muni de la tribu produit. Soit $\varphi : \ (x, \theta) \mapsto x$ la projection sur $\boldsymbol{X}$. On définit une mesure jointe λ sur l'espace produit par*

$$\lambda(A \times B) = \int_B \mathrm{P}_\theta(A) \, \mathrm{d}\pi(\theta) \, , \quad A \in \mathcal{X}, \, B \in \mathcal{T} \, . \tag{1.24}$$

Soit $\{\lambda_x\}_x$ une désintégration de λ par rapport à la projection φ, si elle existe. La mesure λ_x, concentrée sur $\{x\} \times \boldsymbol{\Theta}$ pour tout $x \in \boldsymbol{X}$, est identifiée à une mesure π_x sur $\boldsymbol{\Theta}$, qu'on appelle loi a posteriori *sachant x.*

La mesure $\mu_{\boldsymbol{X}}$, qui est σ-finie d'après la définition 4 est alors la mesure image $\lambda \circ \varphi^{-1}$. En ce sens, on peut dire que π_x est la désintégration de la mesure jointe λ par rapport à la prédictive a priori $\mu_{\boldsymbol{X}}$.

Dénominateur commun des deux approches Que le prior soit propre ou impropre, la loi a posteriori est définie, $\mu_{\boldsymbol{X}}$ presque sûrement, comme une mesure de probabilité (dans le cas du prior impropre, grâce à la proposition 6). La définition donnée dans le cas général coïncide avec celle donnée dans le cas particulier du prior propre. En effet, la mesure λ est alors mesure de probabilité sur $\boldsymbol{X} \times \boldsymbol{\Theta}$ et λ_x en est une probabilité conditionnelle sachant x. Cette probabilité résume l'information disponible après l'observation $X = x$, compte tenu du prior.

Jusqu'à présent, on a seulement donné une définition de la loi a posteriori et rien ne garantit son existence, que le prior soit propre ou non. Le paragraphe suivant règle la question et donne l'expression de la densité du posterior par rapport au prior.

Formule de Bayes pour la loi a posteriori Pour éviter dans un premier temps une trop grande technicité, on fait l'hypothèse suivante.

Hypothèse 1.
L'application $p_x : \theta \mapsto p_\theta(x)$ est mesurable pour ν-presque tout x.

Notons que cette hypothèse n'est pas automatiquement satisfaite car la mesurabilité de l'application $\theta \mapsto \mathrm{P}_\theta(A)$ pour tout A n'entraîne pas celle de p_x. Il est toutefois prouvé en appendice A que l'hypothèse 1 est vérifiée sous une autre hypothèse très générale de régularité de l'espace des observations.

Remarque 2. D'après la définition de la prédictive a priori $\mu_{\boldsymbol{X}}$, la mesure de référence ν domine $\mu_{\boldsymbol{X}}$. L'hypothèse 1 assure de plus, en utilisant le théorème de Tonelli-Fubini, que la densité de $\mu_{\boldsymbol{X}}$ par rapport à ν est

$$\frac{\mathrm{d}\mu_{\boldsymbol{X}}}{\mathrm{d}\nu}(x) = \int_{\boldsymbol{\Theta}} p_\theta(x)\,\mathrm{d}\pi(\theta) \ := m(x)\,.$$

Ainsi, la densité marginale m est $\mu_{\boldsymbol{X}}$-presque sûrement non nulle, et finie, puisque $\mu_{\boldsymbol{X}}$ est σ-finie.

Théorème 7 (formule de Bayes)**.** *La loi a posteriori sachant l'observation $X = x$ existe pour $\mu_{\boldsymbol{X}}$-presque tout x. La densité relativement au prior π est donnée, $\mu_{\boldsymbol{X}}$-presque partout, par*

$$\frac{\mathrm{d}\pi_x}{\mathrm{d}\pi}(\theta) = \frac{p_\theta(x)}{\int_{\boldsymbol{\Theta}} p_t(x)\,\mathrm{d}\pi(t)} \tag{1.25}$$

Démonstration. La preuve suit en grande partie Schervish (1995). D'après la remarque 2, il existe un ensemble $X_\mu \subset \boldsymbol{X}$, avec $\mu_{\boldsymbol{X}}(X_\mu) = 1$, tel que pour tout $x \in X_\mu$, $0 < m(x) < \infty$, de sorte que le dénominateur du membre de droite dans (1.25) est $\mu_{\boldsymbol{X}}$-presque partout fini et non nul. Par l'hypothèse 1, on peut choisir X_μ de sorte que l'application partielle p_x soit $\mathcal{T}$-mesurable pour $x \in X_\mu$. Ainsi, le membre de droite définit bien une densité sur $\mathcal{T}$. Pour $x \in X_\mu$, soit $\tilde{\pi}_x$ la mesure dont la densité par rapport à π est donnée par le membre de droite. On va montrer que $\tilde{\pi}_x$ est une version de la la loi a posteriori, au sens de la définition générale 7.

Il suffit de montrer que, pour toute fonction g, λ-mesurable, positive, le point (ii) de la définition 6 est vérifié avec $\lambda_x = \mathbb{1}_{\{x\}} \times \tilde{\pi}_x$ et $\mu = \mu_{\boldsymbol{X}}$. Or, pour une telle fonction g,

$$\begin{aligned}
\lambda_x.g &:= \int_{\boldsymbol{X}\times\boldsymbol{\Theta}} g(s,\theta)\,\mathrm{d}\left[\mathbb{1}_{\{x\}} \times \tilde{\pi}_x\right](s,\theta) \\
&= \int_{\boldsymbol{\Theta}} g(x,\theta)\,\mathrm{d}\tilde{\pi}_x(\theta) \\
&= \int_{\boldsymbol{\Theta}} g(x,\theta)\frac{\mathrm{d}\tilde{\pi}_x}{\mathrm{d}\pi}\,\mathrm{d}\pi(\theta) \\
&= \frac{1}{m(x)}\int_{\boldsymbol{\Theta}} g(x,\theta)p_\theta(x)\,\mathrm{d}\pi(\theta)\,.
\end{aligned}$$

La fonction m est mesurable et $\mu_{\boldsymbol{X}}$-presque partout non nulle, car elle est la densité $\frac{\mathrm{d}\mu_{\boldsymbol{X}}}{\mathrm{d}\nu}$ (*c.f.* la remarque 2). De plus, puisque g est λ-intégrable, avec

$$\lambda.g := \int_{\boldsymbol{X}\times\boldsymbol{\Theta}} g(x,\theta)p_\theta(x)\,\mathrm{d}\nu(x)\,\mathrm{d}\pi(\theta)\,,$$

La fonction $(x,\theta) \mapsto g(x,\theta)p_\theta(x)$ est $\nu\times\pi$-intégrable et le théorème de Tonelli assure la mesurabilité de $x \mapsto \int_{\boldsymbol{\Theta}} g(x,\theta)p_\theta(x)\,\mathrm{d}\pi(\theta)$. Ainsi, l'application $x \mapsto \lambda_x.g$ est $\mathcal{X}$-mesurable et l'on a

$$\begin{aligned}
\int_{\boldsymbol{X}} \lambda_x.g\,\mathrm{d}\mu_{\boldsymbol{X}}(x) &= \int_{\boldsymbol{X}}\int_{\boldsymbol{\Theta}} g(x,\theta)\frac{p_\theta(x)}{m(x)}\,\mathrm{d}\pi(\theta)\,\mathrm{d}\mu_x(x) \\
&= \int_{\boldsymbol{X}\times\boldsymbol{\Theta}} g(x,\theta)\frac{p_\theta(x)}{m(x)}m(x)\,\mathrm{d}\pi(\theta)\,\mathrm{d}\nu(x) \\
&\overset{\text{Fubini}}{=} \int_{\boldsymbol{\Theta}}\int_{X} g(x,\theta)p_\theta(x)\,\mathrm{d}\nu(x)\,\mathrm{d}\pi(\theta) \\
&= \lambda.g\,,
\end{aligned}$$

ce qui prouve le point (ii) de la définition 6. La famille $\{\lambda_x\}$ est bien une désintégration de λ par $\mu_{\boldsymbol{X}}$, donc la loi a posteriori π_x existe et l'on peut définir $\frac{\mathrm{d}\pi_x}{\mathrm{d}\pi}$ comme dans (1.25).

□

Remarquons que, pour $x \in X_\mu$, on a directement, sans utiliser la proposition 6, le fait que $\pi_x(\boldsymbol{\Theta}) = \int_{\boldsymbol{\Theta}} \frac{\mathrm{d}\pi_x}{\mathrm{d}\pi}\,\mathrm{d}\pi = 1$, donc que π_x est une mesure de probabilité.

Dans un contexte statistique, l'observation consiste en un échantillon $X_{1:n} = x_{1:n}$ et l'on suppose les variables X_i indépendantes et identiquement distribuées (*i.i.d.*) selon une certaine loi $\mathrm{P}_\theta \in \mathcal{P}_{\boldsymbol{\Theta}}$. La définition de la loi a posteriori est alors naturellement modifiée en remplaçant $p_\theta(x)$ par $\prod_i p_\theta(x_i)$, cette dernière quantité étant bien la densité de la loi produit $P_\theta^{\otimes n}$ par rapport à la mesure de référence $\nu^{\otimes n}$ sur $\mathcal{X}^{\otimes n}$.

1.3.2 Théorie bayésienne de la décision

L'objectif de cette section est d'expliquer le recours quasiment systématique dans cette thèse aux estimateurs ponctuels construits à partir de l'espérance a posteriori, ainsi qu'à des intervalles de « crédibilité » bayésiens, définis comme des inter-quantiles a posteriori, au lieu des intervalles de confiance classiques. Les notions introduites ci-dessous sont présentées par exemple au chapitre 4 de Robert (2007).

Dans la suite, $(\boldsymbol{D}, \mathcal{D})$ est l'espace des décisions et une règle de décision δ est une statistique à valeurs dans $\boldsymbol{D}$, c'est-à-dire une fonction mesurable des observations. On note $\boldsymbol{\Delta}$ l'ensemble de telles règles. Dans cette section, on ne précise pas la taille de l'échantillon dans les notations, de sorte que X (*resp.* x) désigne indifféremment une variable aléatoire (*resp.* une réalisation), ou une suite *i.i.d.* $X_1, \ldots, X_n$ de taille finie (*resp.* un échantillon $x_1, \ldots, x_n$).

On suppose l'existence d'une fonction de coût $L : \boldsymbol{\Theta} \times \boldsymbol{D} \to \mathbf{R}^+$, mesurable en ses deux variables. L'existence d'une telle fonction dans des situations pourtant assez communes n'est pas assurée, mais nous ne traiterons pas ce problème ici. Pour une discussion de tels paradoxes et une formulation rigoureuse des axiomes de la théorie de la décision, voir par exemple Robert (2007), chapitre 2. Pour alléger les notations, on note $\mathrm{d}\nu(x) = \mathrm{d}x$ lorsqu'il n'y a pas d'ambiguïté.

Définition 8 (Risque classique)**.** *Le risque $R(\theta, \delta)$ associé à la règle de décision $\delta \in \boldsymbol{\Delta}$ est l'espérance, sur toutes les réalisations possibles de X à θ fixé, de la fonction de coût,*

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta\left[L(\theta, \delta(X))\right] \\ &= \int_{\boldsymbol{X}} L(\theta, \delta(x)) p_\theta(x)\,\mathrm{d}x\,. \end{aligned}$$

Le risque classique dépend donc de θ et ne définit qu'un ordre partiel sur $\boldsymbol{\Delta}$, $\delta_1 \succeq \delta_2$ si $\forall\theta$, $R(\theta, \delta_1) \leq R(\theta, \delta_2)$. Les règles de décisions *admissibles* en statistique classique sont les éléments de $\boldsymbol{\Delta}$ maximaux pour le pré-ordre $\succ$, c'est-à-dire les règles δ_* telles qu'il n'existe aucun $\delta \in \boldsymbol{\Delta}$, tel que $\forall\theta, R(\theta, \delta) \leq R(\theta, \delta_*)$ avec inégalité stricte pour au moins un θ.

Définition 9. *Le* risque a posteriori*, étant donnée l'observation $X = x$, est*

$$\begin{aligned}\rho^\pi(x,\delta) &= \mathbb{E}_x^\pi\left[L\left(\theta,\delta(x)\right)\right] \\ &= \int_{\boldsymbol{\Theta}} L(\theta,\delta(x))\,\mathrm{d}\pi(\theta|x)\,.\end{aligned}$$

La différence fondamentale (et représentative des différences générales entre approche bayésienne et fréquentiste) entre les deux définitions concerne l'espace sur lequel a lieu l'intégration. Classiquement, on intègre en x, à θ fixé, alors qu'on intègre en θ, conditionnellement à x, dans l'approche bayésienne. Une manière de résoudre la difficulté mentionnée concernant le risque classique consiste à l'intégrer sur le modèle, ce qui est possible dans un cadre bayésien.

Définition 10 (risque intégré)**.** *Le risque intégré, associé à une règle de décsion δ et à un prior π, est défini par*

$$\begin{aligned}r^\pi(\delta) &= \mathbb{E}^\pi\left[R(\theta,\delta)\right] \\ &= \int_{\boldsymbol{\Theta}} R(\theta,\delta)\,\mathrm{d}\pi(\theta)\,.\end{aligned}$$

Le risque intégré, à prior donné, définit bien un ordre total sur les règles de décision. La notion d'estimateur de Bayes suit :

Définition 11 (Règle et risque de Bayes)**.** *Une règle de Bayes est toute règle de décision δ^* minimisant le risque intégré $r^\pi(\,\cdot\,)$. Si une telle règle existe, le risque de Bayes est le risque associé $r^\pi(\delta^*)$.*

Un estimateur admissible correspondant à un risque intégré fini est donc un estimateur de Bayes. Choisir une règle en fonction de ce critère semble, à première vue, contredire le principe de vraisemblance – n'utiliser que les données effectivement observées pour l'inférence. La proposition suivante montre qu'il n'en est rien.

Proposition 7 (Robert (2007), théorème 2.3.2)**.** *Le risque intégré s'obtient par intégration du risque a posteriori sous la prédictive a priori,*

$$r^\pi(\delta) = \int_{\boldsymbol{X}} \rho^\pi(\delta|x)\,\mathrm{d}\mu_{\boldsymbol{X}}(x)\,,$$

de sorte que toute règle minimisant le risque a posteriori, sachant x, pour tout x, est une règle de Bayes.

Preuve. Le coût étant une fonction positive, on peut utiliser le théorème de Fubini, de sorte que

$$\begin{aligned}r^\pi(\delta) &= \int_{\boldsymbol{X}}\left(\int_{\boldsymbol{\Theta}} L(\theta,\delta(x))p_\theta(x)\,\mathrm{d}\pi(\theta)\right)\mathrm{d}x\,, \\ &= \int_{\boldsymbol{X}}\left(\int_{\boldsymbol{\Theta}} L(\theta,\delta(x))\frac{\mathrm{d}\pi_x}{\mathrm{d}\pi}(\theta)\frac{\mathrm{d}\mu_{\boldsymbol{X}}}{\mathrm{d}x}\,\mathrm{d}\pi(\theta)\right)\mathrm{d}x\,, \\ &= \int_{\boldsymbol{X}}\left(\int_{\boldsymbol{\Theta}} L(\theta,\delta(x))\,\mathrm{d}\pi_x(\theta))\right)\mathrm{d}\mu_{\boldsymbol{X}}(x)\,.\end{aligned}$$

□

Coût quadratique et prédictive a posteriori Supposons qu'on cherche à estimer une quantité réelle $g(\theta)$ dépendant du paramètre. On a alors $\boldsymbol{D} = \mathbf{R}$ et les règles de décision sont des estimateurs $\delta(x) = \hat{g}_x$. En optant pour le coût quadratique

$$L(\theta, \hat{g}_x) = (g(\theta) - \hat{g}_x)^2 \ , \tag{1.26}$$

on a une règle de Bayes de manière explicite.

Proposition 8 (Robert (2007), proposition 2.5.1)**.** *Étant donné un prior π et le coût quadratique* (1.26)*, l'estimateur $\delta(x) = \hat{g}_x$ correspondant à l'espérance a posteriori,*

$$\begin{aligned} \hat{g}_x &:= \mathbb{E}^\pi(g(\theta)|x) \\ &= \int_\theta g(\theta)\, \mathrm{d}\pi(\theta|x) \ , \end{aligned} \tag{1.27}$$

est un estimateur de Bayes.

Preuve. Soit δ_x une estimation de g, à x donné. Le risque a posteriori est

$$\begin{aligned} \rho^\pi(\delta_x, x)) &= \int_{\boldsymbol{\Theta}} \left(g^2(\theta) - 2\,\delta_x g(\theta) + \delta_x^2\right)\, \mathrm{d}\pi_x(\theta) \\ &= \mathbb{E}_x^\pi(g^2) - 2\, g_x \mathbb{E}_x^\pi(g) + \delta_x^2 \, . \end{aligned}$$

Le minimum de cette quantité, en tant que fonction de δ_x, est atteint en $\delta_x = \mathbb{E}_x^\pi(g)$. Le résultat suit à l'aide de la proposition 7. □

En particulier, dans un cadre d'estimation de densité (*resp.* de loi), on définit la *prédictive a posteriori* comme l'espérance a posteriori des densités (*resp.* lois) du modèle :

$$\begin{gathered} \forall y \in \boldsymbol{X}, \quad \hat{f}_x(y) = \int_{\boldsymbol{\Theta}} f(y, \theta)\, \mathrm{d}\pi_x(\theta) \\ \left(resp. \forall A \in \mathcal{X}, \quad \hat{\mathrm{P}}_x(A) = \int_{\boldsymbol{\Theta}} \mathrm{P}_\theta(A)\, \mathrm{d}\pi_x(\theta) \right) . \end{gathered} \tag{1.28}$$

Ainsi définie, la prédictive a posteriori est un estimateur de Bayes pour les fonction de coût $L^y(\theta, \delta) = (f(y, \theta) - \delta)^2$, $y \in \boldsymbol{X}$ (*resp.* $L^A(\theta, \delta) = (P_\theta(A) - \delta)^2$, $A \in \mathcal{X}$). Dans la suite de cette thèse, les prédictives a posteriori des mesures angulaires H définies en section 1.2.2 seront régulièrement utilisées pour évaluer la qualité de l'ajustement d'un modèle d'extrêmes multivariés.

Régions de crédibilité En pratique, être capable de donner un intervalle représentant le niveau de confiance associé à un estimateur, même pour un échantillon de petite taille, est la principale valeur ajoutée de l'inférence bayésienne. D'un point de vue plus théorique, un intérêt des intervalles de

crédibilité bayésiens est de donner un sens direct à l'expression $\mathbb{P}(\theta \in I) = \alpha$, sans avoir à la traduire par $\forall \theta, P_\theta(I \ni \theta) = \alpha$. Lorsque le paramètre n'est pas de dimension finie, on peut s'intéresser, comme au paragraphe précédent, à une quantité réelle $g(\theta)$ dépendant mesurablement de θ, typiquement la densité en un point.

Définition 12. *On appelle* région de crédibilité de niveau $\alpha \in [0, 1]$ *pour* g *toute partie mesurable* $C_\alpha(x) \subset g(\mathbf{\Theta})$ *telle que*

$$\pi \circ g^{-1}(C_\alpha(x)|x) \geq 1 - \alpha\,.$$

De telles régions sont ainsi aléatoires et définies conditionnellement à l'observation $X = x$. Il n'y a bien sûr pas unicité dans la définition précédente. Étant donnée une mesure de référence, on peut par exemple définir des régions de crédibilité de plus forte densité a posteriori, dont le volume est minimal parmi les régions de crédibilité de niveau donné. Dans cette thèse, on choisira par commodité les régions inter-quantiles de niveau $\alpha/2, 1 - \alpha/2$, pour la loi a posteriori $\pi \circ g^{-1}$. De telles quantiles sont en effet aisément accessibles, étant donné un échantillon a posteriori obtenu par exemple par MCMC (*c.f.* section 1.3.4).

1.3.3 Propriétés asymptotiques

Construire des estimateurs assortis de régions de crédibilité satisfaisant les principes de la théorie de la décision n'est pas suffisant ; encore faut-il que l'on soit assuré de leur convergence vers une « vraie valeur » dans la limite des échantillons de grande taille. Lorsque l'espace des paramètres est de dimension finie, sous les conditions de régularité du modèle assurant la consistance de l'estimateur de maximum de vraisemblance, et supposant la positivité de la densité de la loi a priori autour du vrai paramètre, le théorème de Bernstein - von Mises (Van der Vaart, 2000, chapitre 10), assure la normalité asymptotique de la loi a posteriori ; l'estimateur « espérance a posteriori » est asymptotiquement équivalent à l'estimateur de maximum de vraisemblance. Toujours dans un cadre paramétrique, si le modèle est mal spécifié, Berk (1966) montre que, sous des hypothèses de régularité du modèle, la loi a posteriori se concentre sur des régions « asymptotiquement porteuse » (*asymptotic carrier regions*), qui sont les points du modèle minimisant la divergence de Kullback–Leibler par rapport à θ_0. Dans le chapitre 2 de cette thèse, le modèle est une union finie de modèles paramétriques et la consistance est considérée comme acquise.

Dans un cadre non paramétrique, la consistance n'est pas automatique. Tout d'abord, elle dépend de la topologie choisie sur le modèle. Pour éviter les problèmes d'identifiabilité, on prend $\mathbf{\Theta} = \mathcal{P}$.

Définition 13 (Consistance d'un modèle bayésien)**.** *Soit* $\mathscr{T}$ *une topologie sur* $\mathcal{P}$*, et* $\mathcal{T}$ *la tribu borélienne sur* $\mathcal{P}$ *associée. Soit* $\mathrm{P}_0 \in \mathcal{P}$ *la « vraie »*

distribution et $X_i \overset{i.i.d.}{\sim} \mathrm{P}_0$, $i \in \mathbf{N}$. *Soit* π *un prior sur* $\mathcal{T}$. *Le posterior est dit* consistant en P_0 *si,* P_0*-presque sûrement, pour tout voisinage* U *de* P_0,

$$\pi(U|X_{1:n}) \xrightarrow[n\to\infty]{} 1\,. \tag{1.29}$$

Remarque 3. Notons que « P_0-presque sûrement, pour tout U » signifie que l'ensemble négligeable en dehors duquel la convergence a lieu ne dépend pas du choix de U.

Remarque 4. La consistance, à P_0 donnée, dépendant exclusivement des caractéristiques du prior, mais décrivant une propriété du posterior, on parle indifféremment de consistance du prior ou du posterior.

Au chapitre 3 est étudiée la consistance du modèle de mélange de Dirichlet, pour la topologie faible. Il est montré que le modèle est faiblement consistant sous des conditions peu restrictives. La preuve est principalement basée sur un des théorèmes fondamentaux de la théorie asymptotique bayésienne, dû à Schwartz (1965).

1.3.4 Échantillonnage de la loi a posteriori

Dans la plupart des situations concrètes, on définit un prior par sa densité, également notée π, relativement à une mesure de référence $\mathrm{d}\theta$. La définition de la loi a posteriori devient

$$\pi(\theta|x) = \frac{p_\theta(x)\pi(\theta)}{\int_{\boldsymbol{\Theta}} f(x,t)\pi(t)\,\mathrm{d}t}\,. \tag{1.30}$$

L'intégrale au dénominateur n'a bien souvent pas d'expression explicite et la densité a posteriori n'est connue qu'à une constante multiplicative près. Une exception notable est celle des priors conjugués, *i.e.* appartenant à une classe paramétrique stable par conditionnement. Une telle famille n'existe pas nécessairement en dehors des modèles exponentiels. Quand bien même, il peut être plus avantageux de disposer d'un échantillon distribué selon la loi a posteriori que d'une expression explicite de la densité en tout point. En effet, un échantillon permet d'approcher l'espérance et les quantiles a posteriori de n'importe quelle fonctionnelle du paramètre. Il existe un certain nombre de méthodes, dites méthodes de Monte-Carlo, permettant d'échantillonner une loi dont on ne connaît la densité qu'à une constante multiplicative près (Robert et Casella, 2004, 2010). Les algorithmes proposés dans cette thèse sont de type MCMC (Markov Chain Monte-Carlo), construits comme des variantes de l'algorithme de Metropolis-Hastings (Hastings, 1970). Les preuves de convergence sont basées sur des résultats généraux d'ergodicité des chaînes de Markov (CM) (Meyn *et al.*, 1993; Roberts et Rosenthal, 2004, 2006). En effet, le principe général des méthodes MCMC consiste à construire une chaîne de Markov sur l'espace des paramètres de sorte qu'elle « converge » vers sa distribution stationnaire, dans un sens à préciser.

Ergodicité dans les chaînes de Markov Une CM à temps discret sur un espace d'état $(\boldsymbol{\Theta}, \mathcal{T})$ est une suite de v.a. $(\Theta_i)_{i\in\mathbf{N}}$ telle que $\forall i \geq 0$, conditionnellement à Θ_i, Θ_{i+1} soit indépendante de $(\{\Theta_j\}_{j<i}$ (propriété de Markov), de sorte que l'état courant Θ_i ne dépend que de l'état immédiatement précédent. Plus formellement, étant donnée une filtration $\mathcal{F} = (\mathcal{F}_t)_t$ sur Ω, une CM est une suite de v.a. $\mathcal{F}_t$-adaptées, telles que, en notant respectivement $\perp\!\!\!\perp$ et $\perp\!\!\!\perp_\Theta$ l'indépendance et l'indépendance conditionnellement à une variable Θ, on ait

$$\Theta_{i+1} \perp\!\!\!\perp_{\Theta_i} \mathcal{F}_i \,.$$

Une chaîne est dite *homogène* si la loi $[\Theta_{i+1}|\Theta_i]$ ne dépend pas de i. Une telle chaîne est entièrement définie (en loi) par un état initial θ_0 et un noyau de transition $K_\theta(\,\cdot\,)$ (la loi $[\Theta_{i+1}|\Theta_i = \theta]$). Soit Π_n la loi de l'état Θ_n. Elle est le résultat de n applications successives du noyau de transition à l'état initial,

$$K^n_{\theta_0}(\,\cdot\,) = \underbrace{K \star \cdots \star K}_{n-1 \text{ fois}} \star K_{\theta_0}(\,\cdot\,)\,,$$

où $K \star \Pi$ est la mesure Π « décalée » (*shifted*) par le noyau K,

$$K \star \Pi(\,\cdot\,) = \int_{\boldsymbol{\Theta}} K_t(\,\cdot\,)\,\mathrm{d}\Pi(t)\,. \tag{1.31}$$

Dans le cadre bayésien qui nous occupe, on cherche à échantillonner $\boldsymbol{\Theta}$ « approximativement » selon une loi cible $\tilde{\Pi}$. La définition suivante précise la notion d'approximation employée de manière classique, et en particulier au chapitre 3.

Définition 14 (Convergence en variation totale)**.** *La distance de variation totale entre deux mesures de probabilité* Π_1, Π_2 *sur* $\boldsymbol{\Theta}$ *est*

$$d_{\mathrm{TV}}(\Pi_1, \Pi_2) = \sup_{A\in\mathcal{T}} \{\Pi_1(A) - \Pi_2(A)\}\,.$$

On dit qu'une CM $(\Theta_n)_n$ *d'état initial* θ_0 *sur* $\boldsymbol{\Theta}$ converge en variation totale *vers* $\tilde{\Pi}$ *si*

$$d_{\mathrm{TV}}(K^n_{\theta_0}, \tilde{\Pi}) \xrightarrow[n\to\infty]{} 0\,.$$

Une implication très utile en pratique de la convergence en variation totale est l'existence d'une loi des grands nombres

$$\frac{1}{n}\sum_{i=1} g(\Theta_i) \xrightarrow[n\to\infty]{p.s.} \int g\,\mathrm{d}\tilde{\Pi}\,, \tag{1.32}$$

pour toute fonction g, $\tilde{\Pi}$-intégrable.

Définition 15 (Mesure invariante)**.** *La mesure* $\tilde{\Pi}$ *est dite* invariante *par* K *(ou encore* stationnaire*) si*

$$K \star \tilde{\Pi} = K\,.$$

Définition 16 (Cycles et apériodicité). *Un p-cycle est une partition finie $A_1, \ldots, A_p$ de $\mathbf{\Theta}$ telle que*

$$\forall i < p,\ \forall \theta \in A_i\,,\ K_\theta(A_{i+1}) = 1\,, \qquad \text{et} \qquad \forall \theta \in A_p\,,\ K_\theta(A_1) = 1\,.$$

Une CM *est* périodique *de période p si elle admet un p-cycle. S'il n'existe pas de p-cycle pour $p > 1$, elle est dite* apériodique.

Définition 17 (ϕ-irréductibilité). *Une* CM *est dite* ϕ-irréductible *s'il existe une mesure η sur $\mathbf{\Theta}$, σ-finie, telle que $\forall A \in \mathcal{T}$ tel que $\eta(A) > 0$,*

$$\forall \theta \in \mathbf{\Theta},\ \exists n, \quad K_\theta^n(A) > 0\,.$$

Une condition équivalente est que le temps d'attente τ_A avant la première entrée dans A soit presque sûrement fini.

Théorème 8 (Roberts et Rosenthal (2004), théorème 4). *Si une chaîne de Markov est apériodique, ϕ-irréductible et admet $\tilde{\Pi}$ comme distribution stationnaire, elle converge en variation totale vers $\tilde{\Pi}$ pour $\tilde{\Pi}$-presque tout état initial θ_0.*

Sous une condition supplémentaire dite de *Harris-récurrence*, la convergence a lieu pour toute valeur initiale. Cette condition est toutefois difficile à vérifier en général, en particulier dans le cadre des mélanges de Dirichlet.

Les méthodes MCMC en général consistent à construire une CM vérifiant les conditions du théorème 8 avec $\tilde{\Pi} = \pi(\,\cdot\,|x)$ comme loi cible.

Algorithme de Metropolis-Hastings Cet algorithme est considéré comme le plus simple de la classe des MCMC. Une façon d'assurer la stationnarité de la loi cible consiste à imposer une condition de *réversibilité*,

$$\iint_{s\in A, t\in B} \mathrm{d}\tilde{\Pi}(s)\,\mathrm{d}K_s(t) = \iint_{t\in B, s\in A} \mathrm{d}\tilde{\Pi}(t)\,\mathrm{d}K_t(s)\,, \quad A, B \in \mathcal{T}\,. \tag{1.33}$$

On construit pour cela un noyau mixte : On se donne une loi de proposition $Q(s,\,\cdot\,)$, de densité $q(s,\,\cdot\,)$ par rapport à une certaine mesure de référence sur $\mathbf{\Theta}$, et une probabilité d'acceptation (de la proposition) $\alpha(s,t)$. Rappelons que la densité de loi cible est de type $\frac{\mathrm{d}\tilde{\Pi}}{\mathrm{d}\theta} = \gamma\tilde{\pi}$, γ inconnu, de sorte que le rapport $\tilde{\pi}(t)/\tilde{\pi}(s)$ est connu, pour $s, t \in \mathbf{\Theta}$. L'algorithme est le suivant.

Algorithme 1 (Metropolis-Hastings (MH)). *Générer θ_0 selon une loi quelconque, puis, à l'étape n ($n \geq 1$),*

- *Générer $\theta^* \sim Q(\theta_n,\,\cdot\,)$.*
- *Avec probabilité $\alpha = \alpha(\theta_n, \theta^*)$, accepter θ^* et poser $\theta_{n+1} = \theta^*$, avec*

$$\alpha(s,t) = \min\left\{1, \frac{\tilde{\pi}(t)}{\tilde{\pi}(s)}\frac{q(t,s)}{q(s,t)}\right\}\,. \tag{1.34}$$

- *Avec probabilité* $(1-\alpha)$*, rejeter la proposition et poser* $\theta_{n+1} = \theta_n$*.*
- $n \leftarrow n+1$*, et itérer.*

L'algorithme MH définit un noyau de transition mixte,

$$K_s(A) = r(s)\delta_s(A) + \int_A q(s,t)\alpha(s,t)\,\mathrm{d}t\,, \quad s \in \mathbf{\Theta}\,.$$

où δ_θ désigne la masse de Dirac au point θ, et $r(s)$ est la probabilité intégrée de rejet, soit

$$r(s) = \int_{\mathbf{\Theta}} q(s,t)(1-\alpha(s,t))\,\mathrm{d}t\,.$$

Du fait que

$$\begin{aligned}\int_{A\times B} \mathrm{d}\tilde{\Pi}(s)\,\mathrm{d}K_s(t) &= \gamma \int_{A\times B} \tilde{\pi}(s)q(s,t)\alpha(s,t)\,\mathrm{d}t\,\mathrm{d}s + \gamma \int_{A\cap B} \tilde{\pi}(s)r(s)\,\mathrm{d}s \\ &= \int_{B\times A} \mathrm{d}\tilde{\Pi}(s)\,\mathrm{d}K_s(t) + \cdots \\ &\cdots \gamma \int_{A\times B} \left[\tilde{\pi}(s)q(s,t)\alpha(s,t) - \tilde{\pi}(t)q(t,s)\alpha(t,s)\right]\,\mathrm{d}t\,\mathrm{d}s\,,\end{aligned} \tag{1.35}$$

et que, d'après (1.34),

$$q(s,t)\tilde{\pi}(t)\alpha(s,t) = q(t,s)\tilde{\pi}(s)\alpha(t,s), \quad (s,t) \in \mathbf{\Theta}^2\,,$$

K satisfait (1.33). K admet donc $\tilde{\Pi}$ comme loi stationnaire. Enfin, en choisissant un noyau de proposition Q tel que $\pi \ll Q(t,\,\cdot\,)$ pour tout t, le caractère ϕ-irréductible et l'apériodicité sont immédiats. L'algorithme MH génère donc bien une chaîne qui converge, en variation totale, vers la loi cible.

Variantes de type « Gibbs » En pratique, la vitesse de convergence peut être extrêmement lente si le taux d'acceptation est trop faible, ce qui est le cas lorsque les propositions sont générées trop loin de l'état courant.

Dans le cas où le paramètre vit dans une partie de $\mathbf{R}^k$, une manière de générer des propositions qui ne modifient pas trop l'état courant θ_n est d'autoriser des modifications par blocs de composantes, en écrivant $\theta = (\theta^1, \ldots, \theta^r)$, où la somme des dimensions des θ^i est k. Chaque bloc est mis à jour selon une étape de l'algorithme MH. Ceci définit l'algorithme dit de « Metropolis-within-Gibbs ». Le noyau de transition K s'écrit alors comme un mélange de noyaux partiels $K^1, \ldots, K^r$,

$$K_s(\,\cdot\,) = \omega_m \sum_{m=1}^{r} K_s^{(m)}(\,\cdot\,)\,, \quad \sum_m \omega_m = 1\,, \quad \omega_m \geq 0\,.$$

Chaque noyau partiel $K_s^{(m)}$ est supporté par un sous-espace strict de $\boldsymbol{R}^k$ (un cylindre à base $\{\theta^m\}$), de sorte que la loi de proposition ne peut plus dominer

π. On peut tout de même obtenir une chaîne ϕ-irréductible en définissant des propositions dont le support est suffisamment large sur chaque cylindre (voir Roberts et Smith, 1994, pour plus de détails). Le nom de l'algorithme fait référence à l'algorithme de Gibbs, cas particulier dans lequel la loi de chaque bloc conditionnellement aux autres est connue, de sorte que l'on simule tour à tour chaque bloc selon sa loi « conditionnelle totale ».

Algorithme à sauts réversibles (*Reversible-jump algorithm*, Green (1995)) Dans un modèle de mélange à nombre non borné de composants, du type du modèle de Dirichlet étudié au chapitre 3, on peut écrire $\mathbf{\Theta} = \bigcup_{m=1}^{\infty} \mathbf{\Theta}_m$, où chaque $\mathbf{\Theta}_m$ représente l'espace des densités de mélange à m composants, de dimension finie. Il est alors nécessaire de construire une chaîne qui parcoure potentiellement tout l'espace. L'algorithme proposé par Green (1995) permet de construire une telle chaîne en assurant une condition de réversibilité de type (1.33).

Notons $Q\tilde{\Pi}$ la mesure sur $\mathbf{\Theta}^2$ définie par

$$Q\tilde{\Pi}(A \times B) \int_{A \times B} \mathrm{d}\tilde{\Pi}(s)\, \mathrm{d}Q_s(t)\,,$$

et soit $f(s,t)$ sa densité par rapport à une mesure de référence η, supposée *symétrique* sur $\mathbf{\Theta}^2$. Par exemple, dans l'algorithme MH, $f(s,t) = \tilde{\pi}(s)q(s,t)$ et η est le produit des mesures de référence sur $\mathbf{\Theta}$, $\mathrm{d}\eta(s,t) = \mathrm{d}s\,\mathrm{d}t$. En reprenant (1.35), et en intégrant par rapport à $\mathrm{d}\eta(s,t) = \mathrm{d}\eta(t,s)$, la condition de réversibilité (1.33) est vérifiée dès lors que l'on spécifie $\alpha(s,t)$ de sorte que $f(s,t)\alpha(s,t) = f(t,s)\alpha(t,s)$, par exemple comme dans (1.34). Il suffit donc de construire η et de calculer f.

Supposons pour simplifier les notations que pour $n \in \mathbf{N}$, $\mathbf{\Theta}_n$ soit un ouvert de $\mathbf{R}^n$. Définissons les noyaux de transition partiels $K^{(m,p)}$ et $K^{(p,m)}$ pour un saut réversible $\mathbf{\Theta}_m \leftrightarrow \mathbf{\Theta}_p$, $m < p$. On va donc construire une mesure η symétrique sur $(\mathbf{\Theta}_m \cup \mathbf{\Theta}_p)^2$. On suppose que l'on peut passer de $\mathbf{\Theta}_m$ à $\mathbf{\Theta}_p$ en rajoutant les composantes manquantes aux $\theta_m \in \mathbf{\Theta}_m$, c'est-à-dire que l'on suppose l'existence d'un ouvert « de réversibilité » $U \subset \mathbf{R}^{p-m}$ et d'une bijection régulière $\varphi : \mathbf{\Theta}_m \times U \to \mathbf{\Theta}_p$, telle que, pour $s \in \mathbf{\Theta}_m$, la loi de proposition $Q_s(\,\cdot\,)$ soit concentrée sur le cylindre $\varphi(\{s\} \times U)$. On note $q_s^U(u)$ $(u \in U)$ la densité de la proposition $Q_s(\,\cdot\,)$ par rapport à la mesure de Lebesgue sur U.

On définit alors η afin que $Q_s \ll \eta$, en posant, pour $A \in \mathbf{\Theta}_m, B \in \mathbf{\Theta}_p$,

$$\tilde{\eta}(A \times B) = \tilde{\eta}(B \times A) = \ell\{(s,u) \in A \times U : \varphi(s,u) \in B\}\,,$$

où ℓ est la mesure de Lebesgue sur $\mathbf{R}^p$, et pour A et B quelconques dans $\mathbf{\Theta}_m \cup \mathbf{\Theta}_p$,

$$\eta(A \times B) = \tilde{\eta}(A \cap \mathbf{\Theta}_m \times B \cap \mathbf{\Theta}_p) + \tilde{\eta}(B \cap \mathbf{\Theta}_m \times A \cap \mathbf{\Theta}_p)\,.$$

Ainsi, η est symétrique, et la densité de $Q\tilde{\Pi}$ par rapport à η est, pour $s \in \mathbf{\Theta}_m$ et $t = \varphi(s,u) \in \mathbf{\Theta}_p$,

$$\begin{aligned} f(s,t) &= \gamma \tilde{\pi}(s) q_s^U(u)\,, \\ f(t,s) &= \gamma \tilde{\pi}(t) \left| \frac{\partial \varphi}{\partial (s,u)} \right|_{(s,u)} . \end{aligned}$$

Pour résumer, l'algorithme de Green est le suivant. On se donne un ensemble d'arêtes $\{(m,p) \subset \mathbf{N}^2, m \neq p\}$ correspondant aux sauts $\mathbf{\Theta}_m \leftrightarrow \mathbf{\Theta}_p$, tel que pour tout (m,p), $\mathbf{\Theta}_p$ soit accessible depuis $\mathbf{\Theta}_m$ en un nombre fini de sauts. Pour $m < p$, On note $\varphi_{m,p} : (\mathbf{\Theta}_m \times \mathbf{R}^{p-m}) \to \mathbf{\Theta}_p\,, m < p$, les transformations comme ci-dessus. Pour $\theta \in \mathbf{\Theta}_m$, soient $\omega_m(\theta)$ et $\{\omega_{m,p}(\theta)\}$ les poids affectés respectivement aux transitions (m) (intra-$\mathbf{\Theta}_m$) et (m,p) ($\mathbf{\Theta}_m \to \mathbf{\Theta}_p$), de sorte que $\omega_m + \sum_{p \in \mathbf{N}} \omega_{m,p} = 1$.

Algorithme 2 (Reversible-Jump (RJ), Green, 1995)**.** *Tirer θ_0 selon une loi quelconque sur $\mathbf{\Theta}$, puis, pour $n \geq 1$,*

1. *Soit m tel que $\theta_n \in \mathbf{\Theta}_m$. Tirer un type de transition $\tau = (m)$ ou (m,p) selon les probabilités $\omega_m(\theta_n)$, $\omega_{m,p}(\theta_n)$.*
2. *Si $\tau = (m)$, effectuer une étape de l'algorithme MH sur $\mathbf{\Theta}_m$.*
3. *Si $\tau = (m,p)$, $m \neq p$*
   - *Si $m < p$, tirer $u \sim Q_{\theta_n}(\,\cdot\,)$ sur $U = \mathbf{R}^{p-m}$ et poser $\theta^* = \varphi_{m,p}(\theta_n, u)$,*
   $$\alpha = \min \left\{ \frac{\tilde{\pi}(\theta^*)}{\tilde{\pi}(\theta_n)} q_{\theta_n}^U(u)^{-1} \left| \frac{\partial \varphi_{m,p}}{\partial (s,u)} \right|_{(\theta_n,u)} , \ 1 \right\} ,$$
   - *Si $m > p$, poser $(\theta^*, u) = \varphi_{p,m}^{-1}(\theta_n) \in \mathbf{\Theta}_p \times \mathbf{R}^{m-p}$, et*
   $$\alpha = \min \left\{ \frac{\tilde{\pi}(\theta^*)}{\tilde{\pi}(\theta_n)} q_{\theta^*}^U(u) \left| \frac{\partial \varphi_{p,m}}{\partial (s,u)} \right|^{-1}_{(\theta^*,u)} , \ 1 \right\} .$$
   - *Avec probabilité α, accepter θ^* et poser $\theta_{n+1} = \theta^*$.*
   - *Rejeter θ^* avec probabilité $1 - \alpha$ et poser alors $\theta_{n+1} = \theta_n$;*
4. *$n \leftarrow n + 1$.*

Chapitre 2

Combinaison bayésienne de modèles

Au chapitre précédent, on a vu que les lois jointes admissibles pour les extrêmes, qu'elles soient caractérisées par une loi max-stable multivariée ou une mesure angulaire, ne sont pas restreintes à un modèle paramétrique de dimension finie. Cependant, adopter un cadre paramétrique peut dans certains cas faciliter l'interprétation des résultats de l'inférence. Par exemple, s'il s'agit de comparer les caractéristiques des extrêmes de température sur deux périodes temporelles, considérer l'évolution des paramètres est un moyen simple de résumer numériquement l'information. Aucun modèle paramétrique n'est exhaustif, de sorte que plusieurs experts consultés séparément sur un même jeu de données peuvent légitimement utiliser des modèles différents et aboutir à des conclusions différentes. Se pose alors la question des moyens de prendre en compte les différentes estimations. Une possibilité consiste à choisir le « meilleur » modèle, au vu d'un critère adapté à la situation. Ce problème du choix de modèle est bien connu en statistique et a été amplement discuté. Akaike (1973) est un des premiers à proposer un critère de sélection, communément nommé AIC (*Akaike information criterion*). La sélection se fait en minimisant la divergence de Kullback Leibler entre la loi des observations et son estimation dans le modèle, tout en pénalisant la complexité. D'autres critères existent, par exemple le BIC (*Bayesian Information Criterion*), proposé par Schwarz (1978), basé sur l'approximation du rapport de vraisemblance.

Sélectionner un seul modèle, dans un contexte où les modèles rejetés ne sont pas discriminés de manière flagrante, conduit à négliger l'incertitude attachée au choix de modèle. Une alternative consiste à conserver toutes les estimations et à les moyenner en fonction de la vraisemblance marginale de chaque modèle et de son poids a priori. Cette approche est connue sous le nom de *Bayesian Model Averaging* (BMA), que l'on traduit ici par « combinaison bayésienne de modèles ». Le BMA a été utilisé et largement étudié dans de nombreux contextes, (Hoeting *et al.*, 1999; Madigan et Raftery, 1994; Raftery

et al., 2005). Il est essentiellement différent d'une estimation par modèle de mélange. En effet, le modèle correspondant au BMA est une *union* de modèles et non un *produit*. Formellement, si $\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_k$ sont les modèles en concurrence, un modèle de mélange est l'ensemble des combinaisons convexes $\{\sum_{m=1}^{k} \omega_m \, \mathrm{P}_{\theta_m}(\,\cdot\,), \omega_m \geq 0, \sum \omega_m = 1, \theta_m \in \mathbf{\Theta}_m\}$. Au contraire, le modèle correspondant au BMA est

$$\bigcup_{m=1}^{k} \{\mathrm{P}_{\theta_m}(\,\cdot\,), \theta_m \in \mathbf{\Theta}_m\}\,.$$

Concrètement, dans le modèle de mélange, on suppose qu'il existe une variable cachée $Z \in \{1, \ldots, k\}$, $\mathbb{E}(Z_m) = \omega_m$, indicatrice du sous-modèle de provenance de chaque observation, de sorte que deux observations peuvent provenir de deux sous-modèles distincts. À l'inverse, on suppose dans le BMA que les observations proviennent *toutes* d'un des modèles en concurrence. Le « moyennage » s'effectue au stade de l'inférence : les estimateurs de moyenne a posteriori seront intégrés sur l'union des sous-modèles, de sorte qu'ils s'écriront comme une combinaison convexe d'estimateurs intra-modèles. Le BMA est donc un mélange d'estimateurs plutôt que de modèles, comme le souligne E. I. George, dans la discussion faisant suite à la revue de Hoeting *et al.* (1999) sur la question du BMA. Du point de vue de la théorie de la décision, les estimateurs obtenus par BMA sont des estimateurs de Bayes, définis en 1.3.2, pour le coût quadratique. En effet, ils correspondent aux estimateurs obtenus par moyennage sous la loi a posteriori.

Ce chapitre est consacré à l'adaptation du BMA au contexte des extrêmes multivariés, et plus particulièrement à l'estimation de la loi jointe des extrêmes. La première étape consiste à s'assurer que les estimateurs obtenus par combinaison convexe sont valides. Un contre-exemple : si $\hat{G}_1$ et $\hat{G}_2$ sont estimations de lois max-stables, elles-mêmes max-stable, leur moyenne $\lambda\hat{G}_1 + (1-\lambda)\hat{G}_2$, $0 \leq \lambda \leq 1$, ne l'est plus. En revanche, si l'on considère deux mesures angulaires H_1 et H_2, dont les moments d'ordre un sont imposés par (1.16), toute combinaison convexe vérifie à nouveau (1.16) et constitue bien une mesure angulaire valide. L'article qui suit, publié en 2013 dans la revue *Extremes*, développe cette idée.

BAYESIAN MODEL AVERAGING FOR MULTIVARIATE EXTREMES

A. SABOURIN, P. NAVEAU, AND A.-L. FOUGÈRES

Abstract. The main framework of multivariate extreme value theory is well-known in terms of probability, but inference and model choice remain an active research field. Theoretically, an angular measure on the positive quadrant of the unit sphere can describe the dependence among very high values, but no parametric form can entirely capture it. The practitioner often makes an assertive choice and arbitrarily fits a specific parametric angular measure on the data. Another statistician could come up with another model and a completely different estimate. This leads to the problem of how to merge the two different fitted angular measures. One natural way around this issue is to weigh them according to the marginal model likelihoods. This strategy, the so-called Bayesian Model Averaging (BMA), has been extensively studied in various context, but (to our knowledge) it has never been adapted to angular measures. The main goal of this article is to determine if the BMA approach can offer an added value when analyzing extreme values.

keywords: Bayesian model averaging, multivariate extremes, parametric modelling, spectral measure.

MSC: 62F07 and 62F15 and 62H20 and 62H05 and 62P12

1. Introduction

Assessing the probability of occurrence of joint extreme events has proven to be a major issue for risk management and a complex inferential problem in statistics. To illustrate this point, daily maximum concentrations of three air pollutants, PM10 (Particulate matter), NO (Nitrogen oxide) and NO_2 (Nitrogen dioxide), recorded in Leeds (U.K.) during five winter seasons (1994-1998)[1], are displayed in Figure 1. Visually, asymmetrical relationships seem to be present, the dependence between NO2 and NO may be stronger than that between the two other pairs. For this Leeds data set, at least three different approaches (Cooley et al., 2010; Heffernan and Tawn, 2004; Boldi and Davison, 2007) have already been proposed. Heffernan and Tawn (2004)'s study focuses on conditional distributions, allowing both for asymptotic dependence or independence at extreme levels. On the contrary, Cooley et al. (2010) and Boldi and Davison (2007), under the assumption of asymptotic dependence, characterize the joint distribution of extremes in terms of the so-called *angular measure* (see Section 2 for more details), respectively in a parametric and semi-parametric framework. In this paper, we follow this latter approach, and focus on parametric models. Several such models have already been proposed for the case where the data are dependent at asymptotic levels: see *e.g.* chapter 9 of Beirlant et al. (2004); Tawn (1990) or Coles and Tawn (1991) for the Logistic

Date: Preprint version of the article published in *Extremes*, 2013. Received: 23 May 2012 / Revised: 14 November 2012 / Accepted: 22 November 2012.

[1]Five different air pollutant concentrations (PM10, NO, NO_2, O3, and SO_2) were measured in the city centre of Leeds, see `http://www.airquality.co.uk` for more details. We restrict our analysis to the three most dependent pollutants.

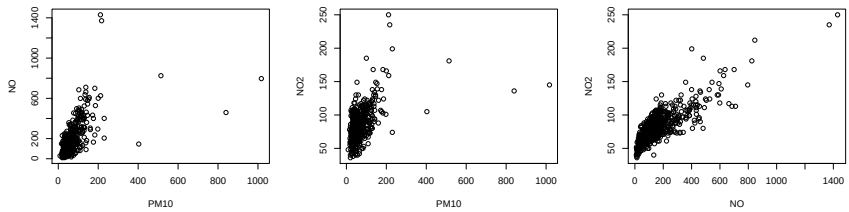

FIGURE 1. Daily maximum concentrations of three air pollutants, PM10, NO and NO_2, recorded in Leeds (U.K.) during five winter seasons (1994-1998)

and Dirichlet families; or Cooley et al. (2010) for the Pairwise Beta model, further generalized by Ballani and Schlather (2011).

In this context, two practitioners working on the same data may well have chosen two distinct models, leading to different estimates of some quantity of interest such as a probability of joint excess of some high multivariate threshold. One could thus reasonably ask if it would be appropriate to merge these results. Our main objective throughout this paper is to investigate how to average the estimates issued from existing parametric families and what are the benefits and the limitations of such an approach.

The Bayesian framework appears to be well tailored for this task because setting priors offers a natural way to integrate results issued from different studies. The so-called *Bayesian Model Averaging* (BMA) method has been extensively studied in other contexts (e.g., Raftery et al., 2005; Madigan and Raftery, 1994; Hoeting et al., 1999). Its adaption to the analysis of multivariate extreme events represents the main aim of this work. To our knowledge, in the field of multivariate extremes, the only publication using BMA is Apputhurai and Stephenson (2011). They combined the cumulative distribution functions of asymptotically dependent and independent models, in the bi-variate case. Our approach differs from theirs in focusing on asymptotically dependent models, and combining the dependence structures themselves (angular measures or exponent functions, see Section 2 for definitions and rationale for such a choice).

In the next section, we recall the necessary background about multivariate extremes. In Section 3, we detail the BMA nuts and bolts within a multivariate extremes context. The BMA scheme is implemented in Section 4 with two different models: the Pairwise Beta model (Cooley et al., 2010) and a nested asymmetric logistic model. A simulation study is performed: data sets are generated from a semi-parametric *Dirichlet mixture model* (DM) introduced by Boldi and Davison (2007) and we compare the predictive performance of the BMA versus a model choice framework. The tri-variate Leeds data set is also revisited. Section 5 offers a few conclusions regarding the advantages and limitations of averaging spectral measures.

2. Background and notations

Spectral measure.
Let $\mathbf{X} = (X_1, \ldots, X_d)^T$ be a positive random vector of dimension d whose margins

follow a unit Fréchet distribution, $\mathbf{P}(X_i \leq x) = \exp(-1/x)$, for all $x > 0$. To describe the extremal behaviour of the vector $\mathbf{X}$, it is mathematically convenient to transform the Cartesian coordinates into pseudo-polar ones by setting

$$R = X_1 + \cdots + X_d \qquad \text{and} \qquad \mathbf{W} = (X_1/R, \ldots, X_d/R)^T$$

where R and $\mathbf{W}$ are often called the radius and the angular vector, respectively. The latter one lies on the unit simplex $\mathbf{S}_d = \{\mathbf{w} : \ w_1 + \cdots + w_d = 1, \quad w_i > 0\}$. With regards to the Leeds data set, we follow the exact same procedure as Cooley et al. (2010) to estimate the marginal distributions of the three pollutants plotted in Figure 1. Each uni-variate series can thus easily be transformed into unit Fréchet distributed ones *via* a probability integral transformation. Observations with the 100 largest radial components [2] (out of 539 non missing triplets) are plotted on the unit simplex $\mathbf{S}_3$ in Figure 2. The points located at the centre of this triangle correspond to events that were equally extreme in the three directions.

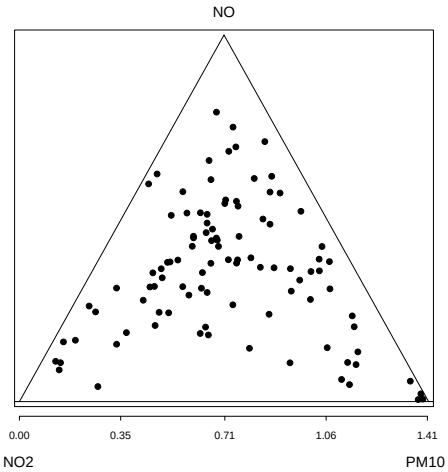

FIGURE 2. Leeds data set: the 100 points with largest radial component $R = x_1 + x_2 + x_3$ (unit Fréchet scale) projected on the unit simplex.

Multivariate extreme value theory tells us that, under mild conditions[3], the dependence structure among excesses above a high radial threshold r can be characterized by the asymptotic distribution H of the angular component:

$$\lim_{r \to \infty} \mathbf{P}\left(\mathbf{W} \in B \mid R > r\right) = H(B)\ , \tag{1}$$

[2]Besides the $\mathbf{L}_1$-norm $(x_1 + x_2 + x_3)$, other norms could be used for threshold selection.

[3] The largest values have to belong to the domain of attraction of a max-stable distribution (the distribution G is said to be max-stable if $G^t(t\mathbf{x}) = G(\mathbf{x})$ for any $t > 0$). This type of distribution arises as the natural non-degenerate limit of rescaled i.i.d. component wise maxima of random vectors with unit Fréchet margins (de Haan and Ferreira, 2006; Resnick, 1987, 2007). Within this framework, it is classical to define the exponent function $V(\mathbf{x}) = -\log G(\mathbf{x})$ that satisfies $V(t\mathbf{x}) = t^{-1}V(\mathbf{x})$.

The *spectral measure* $H(.)$ is any probability measure on the simplex $\mathbf{S}_d$ that satisfies the following moment constraint

$$\forall i \in \{1, \ldots, d\}, \int_{\mathbf{S}_d} w_i \, \mathrm{d}H(\mathbf{w}) = \frac{1}{d} \, . \tag{2}$$

Limit measure.
With our normalization choice, the spectral measure is related to a *limit measure* ν, defined on $\mathbf{E} = [\mathbf{0}, \infty]^d \setminus \{\mathbf{0}\}$, in pseudo-polar coordinates, by $\mathrm{d}\nu = \frac{d}{r^2} \, \mathrm{d}r \, \mathrm{d}H$ (see *e.g.* Chapter 6 of Resnick, 2007). The measure ν is homogeneous of order -1, *i.e.* for any measurable subset $A \subset \mathbf{E}$, $\nu(tA) = \frac{1}{t}\nu(A)$. If A is relatively compact in $\mathbf{E}$,

$$\lim_{n \to \infty} n\mathbf{P}\left(\frac{\mathbf{X}}{n} \in A\right) = \nu(A). \tag{3}$$

In particular, (3) holds for any failure set A of the form $A(\mathbf{u}) = \{\mathbf{x} : \ x_1 > u_1, \ldots, x_d > u_d\}$.

Modelling threshold excesses.
Equations (1) and (3) provide the main elements for modelling excesses in practice. Given a data set whose margins have been transformed into unit Fréchet, one may fix a high radial threshold r_0 and retain only observations with radial component exceeding r_0. The corresponding angular data set $\mathscr{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_n)$, as in Figure 2 with $n = 100$ excesses, is assumed to be an i.i.d. sample distributed according to $H(.)$. Then, the statistician has to propose and fit an adequate spectral measure.

In other words, all the inference in this paper is based on the following key assumption, in view of (3): Conditionally on the radial component R exceeding the retained threshold r_0, the random vector X is assumed to be distributed according to some (normalized) limit probability measure $\tilde{\nu}$, with, in polar coordinates, $\mathrm{d}\tilde{\nu} = \frac{d}{r_0} \frac{\mathrm{d}r}{r^2} \, \mathrm{d}H$. The angular and the radial components are thus assumed to be independent on regions $\{r > r_0\}$, and H characterizes $\tilde{\nu}$, so that a a statistical model for excesses above r_0 can be indexed by a set of angular measures. Also, the likelihood is proportional to the density h evaluated at the angular data points $\mathcal{W}$ and inference can be made with the angular components only. This assumption of 'perfect threshold' has a second consequence: The likelihood of an angular measure which mass is concentrated on the axis is zero, because all the angular data points lie in the interior of the positive quadrant. This restricts any likelihood-based inference to asymptotically dependent models, *i.e.* to H-families which put some mass in the interior of the unit simplex only.

Relaxing the 'perfect threshold' assumption is possible if one works with max-stable distributions, but then, the link with questions related to excesses above threshold is not immediate. Another reason for not considering this option in the context of model averaging is the fact that an average of max-stable distributions is not max-stable. More details are given at the end of this section.

Further, one may consider asymptotically independent models with second-order regular variation in the interior of the positive quadrant (Ledford and Tawn, 1996; Ramos and Ledford, 2009). However, flexible parametric models for asymptotically independent data, in problems of dimension greater than three, have only

recently been proposed in an unpublished paper from Qin *et. al* (2008) [4]. For the sake of simplicity, we leave apart this class of models and focus on asymptotically dependent data.

If a vector $\mathbf{u} = (u_1, \ldots, u_d)^T$ defines the boundary of a failure region $A(\mathbf{u}) = \{\mathbf{x} : \ x_1 > u_1, \ldots, x_d > u_d\}$ is such that $\sum_{i=1}^d u_i > r_0$ for some large r_0, using (3) and the homogeneity property of the limit measure, the probability of being in the failure region can be approximated with ν:

$$\begin{aligned} \mathbf{P}(X_1 > u_1, \ldots, X_d > u_d) \simeq \nu\big(A(\mathbf{u})\big) &= d \int_A \frac{1}{r^2} \, \mathrm{d}r \, \mathrm{d}H(\mathbf{w}) \\ &= d \int_{\mathbf{S}_d} \int_{r > \max_{i=1:d} \frac{u_i}{w_i}} \frac{1}{r^2} \, \mathrm{d}r \, \mathrm{d}H(\mathbf{w}) \\ &= d \int_{\mathbf{S}_d} \min_{i=1:d} \frac{w_i}{u_i} \, \mathrm{d}H(\mathbf{w}) \end{aligned} \tag{4}$$

A classical way of proposing parametric max-stable models is to define them through their *exponent function* V, (see footnote 3), which is related to ν by

$$\forall \mathbf{x} \in \mathbf{E}, \ V(\mathbf{x}) = \nu \left\{ ([0, x_1] \times \cdots \times [0, x_d])^c \right\} ,$$

were $(\,\cdot\,)^c$ denotes the complementary set in $\mathbf{E}$. In the case where all the mass of the angular measure H is concentrated in the interior of the simplex $\mathbf{S}_d$, and V is regular, Theorem 1 of Coles and Tawn (1991) provides a general relationship to derive the spectral density $h(\mathbf{w})$ from $V(.)$: $h(\mathbf{w}) = -\frac{1}{d} \, \partial_{x_1, \ldots, x_d} V(\mathbf{x})|_{\mathbf{x}=\mathbf{w}}$. One advantage of such models is that (4) has an analytical expression obtained from V by inclusion-exclusion. For three-dimensional sample spaces, it yields

$$\begin{aligned} \nu\big(A(\mathbf{u})\big) = V(u_1, u_2, u_3) + \cdots \\ \cdots V(u_1, \infty, \infty) + V(\infty, u_2, \infty) + V(\infty, \infty, u_3) - \ \cdots \\ \cdots (V(u_1, u_2, \infty) + V(u_1, \infty, u_3) + V(\infty, u_2, u_3)) \ . \end{aligned} \tag{5}$$

One drawback is that the angular likelihood h has to be computed by differentiation of order d.

Multivariate extreme models.
In theory, the only constraint on H is encapsulated by (2), which advocates in favour of fully non-parametric estimation methods (see *e.g.* Einmahl et al., 2001; Einmahl and Segers, 2009; Guillotte et al., 2011; Gudendorf and Segers, 2011). In a Bayesian context, it is computationally difficult to handle moderate dimension problems with semi-parametric spectral measures. For example, Boldi and Davison (2007) introduced a semi parametric Bayesian model based on mixtures of Dirichlet distributions and concluded that "one practical drawback with the approach stems from the use of simulation algorithms, which may converge slowly unless they have been tuned. A second is that the number of parameters increases rapidly with the number of mixture components, so model complexity must be sharply penalized through an information criterion or a prior on the number of mixture components".

[4] Qin X., Smith R.L., Ren R.E., Modelling multivariate extreme dependence, In 2008 Joint Statistical Meetings(JSM) Proceedings, Risk Analysis Section. Alexandria, VA: American Statistical Association: 3089-3096

From a practical point of view, parameters in some well-chosen models may offer interpretable summaries to describe the dependence structure (*e.g.*, a finite number of parameters allows to compare between two time periods), and a feasible strategy to reduce the computational complexity. A seminal example (Gumbel, 1960) of parametric model defined by the exponent function is the logistic one

$$V_{\mathrm{L}}(\mathbf{x}) = \left(\sum_{i=1}^{d} x_i^{-1/\alpha} \right)^{\alpha} \quad (0 < \alpha < 1) \ .$$

The logistic model can be extended to handle asymmetrical behaviours and to capture additional dependencies among subsets of variables (Coles and Tawn, 1991). In particular, the exponent function

$$\begin{aligned} V_{\mathrm{NL}}(\mathbf{x}) = 2^{-\alpha_0} \Big[\left(x_1^{\frac{-1}{\alpha_0 \alpha_{12}}} + x_2^{\frac{-1}{\alpha_0 \alpha_{12}}} \right)^{\alpha_{12}} + \left(x_1^{\frac{-1}{\alpha_0 \alpha_{13}}} + x_3^{\frac{-1}{\alpha_0 \alpha_{13}}} \right)^{\alpha_{13}} + \cdots \\ \left(x_2^{\frac{-1}{\alpha_0 \alpha_{23}}} + x_3^{\frac{-1}{\alpha_0 \alpha_{23}}} \right)^{\alpha_{23}} \Big]^{\alpha_0} \quad (0 < \alpha_0 \, , \alpha_{12}, \alpha_{13}, \alpha_{23} < 1) \ , \end{aligned} \tag{6}$$

is a possible generalization which allows for asymmetric pairwise relationships, while concentrating all its mass in the interior of $\mathbf{S}_3$. This model belongs to the class of the *nested asymmetric logistic models*. In the remainder of this paper, we refer to the one defined by (6) as the NL model, we denote $\underline{\alpha} = (\alpha_0, \alpha_{12}, \alpha_{13}, \alpha_{23})$. The expression for the NL density $h_{\mathrm{NL}}(\,\cdot\,|\underline{\alpha})$ on $\mathbf{S}_3$ is given in appendix. One advantage of NL is its low number of parameters and in their interpretability. The scalar α_0 describes the overall dependence among the three coordinates and the α_{ij}'s characterize the additional pairwise dependences. The dependence between a coordinates subset is a decreasing function of the corresponding parameter.

It is also possible to define a multivariate extreme model directly through the spectral density. For example, Cooley et al. (2010) fitted to the Leeds data set the following Pairwise Beta (PB) model

$$h_{\mathrm{PB}}(\mathbf{w}|\beta_0, \{\beta_{ij}\}_{1 \le i < j \le d}) = \sum_{1 \le i < j \le d} h_{ij}(\mathbf{w}|\beta_0, \beta_{ij}) \quad (\beta_0, \beta_{ij} > 0) \ ,$$

which is a sum of beta functions defined by

$$h_{ij}(\mathbf{w}|\beta_0, \beta_{ij}) \quad = \quad K_d(\beta_0) \ w_{ij}^{2\beta_0 - 1} \, (1 - w_{ij})^{(d-2)\beta_0 - d + 2} \, \frac{\Gamma(2\beta_{ij})}{\Gamma^2(\beta_{ij})} w_{i/ij}^{\beta_{ij}-1} w_{j/ij}^{\beta_{ij}-1}$$

with $w_{ij} = w_i + w_j$, $w_{i/ij} = \frac{w_i}{w_i + w_j}$ and $K_d(\beta_0) = \frac{2(d-3)!}{d(d-1)} \frac{\Gamma(\beta_0 d + 1)}{\Gamma(2\beta_0 + 1)\Gamma(\beta_0(d-2))}$. [5] The interpretation for the parameters in the PB model is similar to the NL model's one, except that the strength of the dependence is an increasing function of β_0 and of the β_{ij}'s. Again, we denote $\underline{\beta} = (\beta_0, \beta_{12}, \beta_{13}, \beta_{23})$.

Having at our disposal several spectral models leads us to the question of how to average them with respect to the data set at hand. First, one could wonder what is the meaning of averaging spectral measures in terms of random variables and if directly averaging the corresponding max-stable *distributions* could be a valuable

[5] The difference of a multiplicative factor $\sqrt{d}$ in our normalizing constant compared to the one given by Cooley et al. (2010) is due to the choice of the reference measure: in the aforementioned study, the reference measure is the Lebesgue measure (more precisely the Hausdorff measure) on the simplex itself, whereas we write our densities with respect to its projection on the $d-1$ dimensional euclidean space.

alternative. However, if the random vector M_j follows a max-stable distribution $G_j(\mathbf{x}) = \exp(-V_j(\mathbf{x}))$ with unit Fréchet margins, then the averaged distribution $G(\mathbf{x}) = p_1 G_1(\mathbf{x}) + \cdots + p_J G_J(\mathbf{x})$ (with $\sum_{j=1}^{J} p_j = 1$) is *not* max-stable anymore: it does not satisfy $G^t(t\mathbf{x}) = G(\mathbf{x})$ for any $t > 0$. In contrast, averaging the angular measures H_j still provides another valid angular measure. Indeed, $p_1 H_1(.) + \cdots + p_J H_J(.)$ satisfies (2). In terms of random vectors, averaging angular measures translates into component-wise max-linear combinations. More precisely, if the M_j's are independent, then the max-linear combination $\tilde{M} = p_1 M_1 \vee \cdots \vee p_J M_J$, where $\vee$ denotes the component-wise maximum, has exponent function $\tilde{V}(\mathbf{x}) = p_1 V_1(\mathbf{x}) + \cdots + p_J V_J(\mathbf{x})$. The latter is associated with the spectral measure $\tilde{H} = p_1 H_1 + \cdots + p_J H_J$. This derives immediately from the homogeneity property ($V_j(t\mathbf{x}) = t^{-1} V_j(\mathbf{x})$) characterizing exponent functions:

$$\mathbf{P}(\tilde{M} \leq \mathbf{x}) = \mathbf{P}\left(\bigvee_{j=1}^{J} p_j M_j \leq \mathbf{x}\right) = \prod_{j=1}^{J} \mathbf{P}(M_j \leq \frac{\mathbf{x}}{p_j})$$
$$= \prod_{j=1}^{J} \exp\left[-V_j\left(\frac{\mathbf{x}}{p_j}\right)\right] = \exp\left[-\sum_{j=1}^{J} p_j V_j(\mathbf{x})\right] .$$

3. Bayesian Model Averaging for spectral measures

In the general context of parametric modeling, the information loss relative to the choice of one particular model may be high. Averaging the estimates stemming from several models, with appropriate weights, can be used to partially overcome this issue. As an example, Raftery et al. (2005) found some evidence in an ensemble weather forecast context that the predictive variance in one model would sometimes not reflect the total predictive uncertainty, whereas the predictive variance in the averaged model accounted better for prevision errors. Madigan and Raftery (1994) found some examples, in a contingency tables context, where averaging models resulted in better predictive performance, as measured by a logarithmic scoring rule. We recall here the basic BMA features within our spectral measure context. For a review of BMA, the reader may refer to Hoeting et al. (1999).

Suppose we have M spectral density models $\mathcal{M}_1, \ldots, \mathcal{M}_M$, such that each model $\mathcal{M}_m = \{h_m(\cdot \mid \theta_m),\ \theta_m \in \Theta_m\}$ has a finite dimensional parameter space Θ_m. In this paper, for illustrative purpose, we set $M = 2$ and h_1 and h_2 correspond respectively to the aforementioned PB spectral measure family h_{PB} and to the NL one h_{NL}.

In a Bayesian framework, beliefs of the statistician about θ_m (*e.g.*, arising from expert knowledge), prior to any observation, are made explicit: each parameter space Θ_m is endowed with a *prior* measure (in our case, a probability measure), denoted π_m. Now, in a Bayesian model averaging setting, the statistical model $\tilde{\mathcal{M}}$ is the *disjoint union* of the individual models: in other words, the parameter space $\tilde{\Theta}$ indexing $\tilde{\mathcal{M}}$ is the disjoint union $\tilde{\Theta} = \bigsqcup_{m=1}^{M} \Theta_m$. We recall that a disjoint union of sets $A_1, \ldots, A_M$ is defined by $\bigsqcup_{m=1}^{M} A_i = \{(m, a_m), 1 \leq m \leq M, a_m \in A_m\}$. In the sequel, the term 'union model' will refer to the BMA model indexed by the disjoint union $\tilde{\Theta}$. A prior on the index set $\{1, \ldots, M\}$ is thus needed: we choose $(p_1, \ldots, p_M)$, with $p_1 + \cdots + p_M = 1$, so that p_m is the *a priori* weight of $\mathcal{M}_m$. In

this work, lacking expert knowledge, we choose a uniform prior: $p_m = 1/M$ for all m. The prior distribution $\tilde{\pi}$ on $\tilde{\mathcal{M}}$ is

$$\tilde{\pi}(\bigsqcup_{m=1}^{M} B_m) = \sum_{m=1}^{M} p_m \pi_m(B_m) \ ,$$

for any collection of measurable sets $(B_1, \ldots, B_M)$ with $B_m \subset \Theta_m$. Suppose that each model has a well defined spectral density $h_m(\cdot \mid \theta_m)$. Given the sample of excesses $\mathscr{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_n)$, a common density estimator is the *Posterior predictive density* [6] which, in the disjoint union model, is defined by

$$\tilde{h}(\mathbf{w} \mid \mathscr{W}) = \sum_{m=1}^{M} p_m(\mathscr{W}) \int_{\Theta_m} h_m(\mathbf{w} \mid \theta_m) \, d(\pi_m|\mathscr{W})(\theta_m) \ , \tag{7}$$

where $\pi_m|\mathscr{W} = \pi_m(\cdot \mid \mathscr{W})$ is the posterior distribution restricted to $\mathcal{M}_m$, and $p_m(\mathscr{W})$ is the posterior weight of $\mathcal{M}_m$. This explains the terminology "model averaging": our density estimate is the average of the density estimates produced in separate Bayesian models. As mentioned above, the family of admissible densities is stable under convex combinations, and it is also stable under integration with respect to any probability measure. Consequently, the posterior predictive density still defines a valid angular measure.

More generally, if the goal is to estimate some quantity of interest Δ, (a measurable function of $\underline{\theta}_m$ in each model $\mathcal{M}_m$, such as *e.g.* the probability of a failure set, then Δ is a random variable which prior and posterior distributions in each model are respectively the image measures $\Delta^* \pi_m$ and $\Delta^*[\pi_m|\mathscr{W}]$. The posterior distribution in the BMA model is thus the average

$$\Delta^*[\pi|\mathscr{W}] = \sum_{m=1}^{M} p_m(\mathscr{W}) \Delta^*[\pi_m|\mathscr{W}]$$

and the mean estimate is $\hat{\delta} = \sum_{m=1}^{M} p_m(\mathscr{W}) \int_{\Theta_m} \Delta(\theta_m) \, d[\pi_m|\mathscr{W}](\theta_m)$.

The Bayes formula provides immediately the expression for the posterior weights: $p_m(\mathscr{W})$ is proportional to the *marginal* likelihood $\mathcal{L}_m(\mathscr{W})$ of the observed angular sample in each model $\mathcal{M}_m$, multiplied by the corresponding prior model weight

$$p_m(\mathscr{W}) = \frac{p_m \mathcal{L}_m(\mathscr{W})}{p_1 \mathcal{L}_1(\mathscr{W}) + \cdots + p_M \mathcal{L}_M(\mathscr{W})} \ ,$$

where

$$\mathcal{L}_m(\mathscr{W}) = \int_{\Theta_m} h_m(\mathscr{W}|\theta_m) \, d\pi_m(\theta_m) \ . \tag{8}$$

In practice, for high dimensional parameter spaces, the main hurdle lies in estimating the integral (8). This can be done either by Monte-Carlo methods or asymptotic approximations, from which the Bayesian Information Criterion (BIC) derives. (see *e.g.* the review from Kass and Raftery, 1995, and the references therein). When the sample size n is not too small compared to the dimension k of

[6] The density estimator defined by (7) is a "Bayes estimator": it minimizes the posterior expected quadratic loss $E_{\tilde{\pi}|\mathscr{W}} \left\{ \left(h(\mathbf{w}|\cdot) - \hat{\mathbf{h}}(\mathbf{w}) \right)^2 \right\}$, at each point $\mathbf{w}$, where the expectancy is taken with respect to the *posterior* density $\tilde{\pi}|\mathscr{W}$ in the union parameter space (see e.g. Robert, 2007, Chap. 2, for details about Bayesian decision theory).

the parameter space, a reasonable trade-off between precision and computational efficiency is the Laplace's approximation method: the logarithm of the integrand $\tilde{l}_m(\theta_m) = \log\left[h_m(\mathbf{w}|\theta_m)\pi_m(\theta_m)\right]$ should be approximately normal around the posterior mode $\hat{\theta}_m$, with covariance matrix $\hat{\Sigma} = (-\mathbf{d}^2\tilde{l})^{-1}$ where $\mathbf{d}^2\tilde{l}$ is the Hessian matrix at $\hat{\theta}$. This yields, by integration, the Laplace approximation

$$\hat{\mathcal{L}}_m(\mathscr{W}) = (2\pi)^{k/2}\left|\hat{\Sigma}\right|^{1/2} h(\mathscr{W}|\hat{\theta}_m)\pi_m(\hat{\theta}_m) \tag{9}$$

Kass and Raftery (1995) suggest that in most cases where $n/k \geq 20$, (9) yields a good precision. More details about the validity of (9) may be found in Kass et al. (1990). For lower sample size, one alternative to obtain the posterior weights would be to implement a MC MC algorithm with reversible jumps between the individual models. The proportion of 'time' spent in each model would provide estimates of posterior weights. The main difficulty with this approach would be to define reasonable 'jumps' proposals, to obtain jumps acceptance rates high enough for the chain's mixing properties to remain acceptable in practice.

Inside each single model, the posterior parameter distribution is classically evaluated by a Metropolis-Hastings algorithm producing an approximate posterior sample $(\theta_{m,1}, \ldots, \theta_{m,N})$. The latter is used to approximate each term $\tilde{h}_m(\mathbf{w}) = \int_{\Theta_m} h_m(\mathbf{w} \mid \theta_m) d(\pi_m|\mathscr{W})(\theta_m)$ in (7) *via*

$$\hat{h}_m(\mathbf{w}) = \frac{1}{N}\sum_{t=1}^{N} h_m(\mathbf{w} \mid \theta_{m,t}) \ . \tag{10}$$

Throughout this paper, the total number of simulations is set to 50×10^3, from which the first 30×10^3, values are discarded. The Heidelberger and Welch test (Heidelberger and Welch, 1981; Cowles and Carlin, 1996) and the Geweke convergence diagnostics (Geweke, 1992) show good convergence properties for this burn-in period.

4. BMA with the PB and NL spectral measures

4.1. **Preliminary: definition of Bayesian PB and NL models.** Before implementing the BMA scheme, we need to separately implement our two models in a Bayesian framework. To our knowledge, this has never been done for the PB model neither for our NL model.

Since none of these models is part of the exponential family, there is no obvious uninformative or invariant prior choices at hand. So, for convenience, we transform the parameter spaces to obtain unconstrained ones. Namely, we set

$$\underline{\theta}_{\mathrm{PB}} = \log(\underline{\beta}) \in \mathbf{R}^4\,; \quad \underline{\theta}_{\mathrm{NL}} = \mathrm{logit}(\underline{\alpha}) \in \mathbf{R}^4\,.$$

where, for the NL model, $\mathrm{logit}(x) = \log(x/(1-x))$, which excludes the independent case $\underline{\alpha} = (1,1,1,1)$. Then, the parameters in each model are assumed to be *a priori* mutually independent and normally distributed with common mean equal to 0 and standard deviation equal to 3. Results on simulated data (see Appendix) show that this prior specification does not introduce a strong bias in the estimations.

4.2. **Averaging the PB and NL models: simulation study.**

Comparison with other approaches.
An alternative to the BMA framework would be to select the 'best' model given a data set. The criterion for comparison could be, for example, the posterior weight, or the BIC or AIC. In our case, these three criteria are approximately equivalent: indeed, the differences of scores between two models in terms of AIC or BIC are the same when the two models have same dimension ($k_1 = k_2$), since in such a case $\mathrm{BIC}_{12} - \mathrm{AIC}_{12} = (k_2 - k_1)\log n - 2(k_2 - k_1) = 0$. Selecting the model according to the BIC or the AIC is thus exactly the same. As for the posterior weights, note that the prior model weights were chosen uniform (here $(1/2, 1/2)$). The posterior odds are then equal to the Bayes factor: $B_{12} = p_1(\mathscr{W})/p_2(\mathscr{W}) = \mathcal{L}_1(\mathscr{W})/\mathcal{L}_2(\mathscr{W})$. For large sample sizes the logarithm of the latter can be approximated by the Schwarz criterion $S = \log \mathcal{L}_1(\mathscr{W}) - \log \mathcal{L}_2(\mathscr{W}) - 1/2(k_2 - k_1)\log(n)$, which is $-1/2$ times the the BIC (see *e.g.* Kass and Raftery, 1995; Kass et al., 1990). In view of the asymptotic equivalence of the three criteria, and because posterior weights are anyway computed for the BMA, the 'model selection' alternative considered here consists in selecting the model with highest posterior weight.

The main goal of our simulations is to compare the predictive performance of the BMA against this model selection framework and against single models, in terms of predictive angular densities and estimations of the probability of being in a failure region $A(\mathbf{u})$ as defined in Section 2. The union model is larger than any individual model, and averaging instead of selecting allows to 'integrate' the uncertainty. One could thus expect the predictive performance to be enhanced.

The main theoretical limitation of the averaging approach stems from a concentration phenomenon: if the data arises from the union model (thus, from one of the individual model), the posterior mass should concentrate around the true value and assign more mass to the model containing it. In "misspecified" cases (when the true distribution does not belong to the union model), the posterior is bound to concentrate around "asymptotic carrier" regions of the parameter space, which minimize the Kullback-Leibler divergence to the true distribution (Berk, 1966; Kleijn and van der Vaart, 2006). In our context, this means that, for large sample sizes, the BMA is likely to select a single model, except if the true distribution is at exact "equi-distance" from the two. Consequently, we restrict our study to situations where the sample size is moderate (namely, 80 points).

Predictive scores with simulated data.
In this paragraph, the angular data set $\mathscr{W}$ under consideration is supposed to be simulated according to some angular distribution h_0 (a Dirichlet mixture (DM) distribution, see the next paragraph) on the simplex $\mathbf{S}_3$. The density estimates produced by each inference framework (PB model, NL model, BMA and model selection) are to be compared. We now introduce different scoring rules allowing to do so. The interest of considering several scores is that they rank the predictions according to different criteria. It may happen that one framework be selected by one scoring rule and discarded by another one. The aim here is to check consistency, *i.e.* that our conclusions are relatively independent from the considered score.

As a performance score for a density estimate $\hat{h}$ fitted to $\mathscr{W}$, we consider the logarithmic score

$$LS(\hat{h}, h_0) = -\mathbf{E}_{h_0} \log(\hat{h}(\,\cdot\,)) = -\int_{\mathbf{S}_3} \log(\hat{h}(\mathbf{w}))h_0(\mathbf{w})\,\mathrm{d}\mathbf{w}\,, \tag{11}$$

associated to the Kullback-Leibler divergence between the density estimate and the true distribution. A model with low LS is 'close' to the truth. According to this rule, the best model is the one minimizing the score (note that a zero is not a measure of perfect fit). Since one can simulate from h_0, the latter integral can be evaluated by simple Monte-Carlo

$$\hat{LS}(\hat{h}, h_0) = \frac{-1}{N_{\mathrm{mc}}} \sum_{N=1}^{N_{\mathrm{mc}}} \log(\hat{h}(\mathbf{w}_N))\;; \quad \mathbf{w}_N \overset{i.i.d.}{\sim} h_0\,. \tag{12}$$

In the remainder of this paper, N_{mc} is set to 50×10^3.

The approximation (12) allows us to compare the performance of the predictive densities $\hat{h}_{\mathrm{PB}}, \hat{h}_{\mathrm{NL}}, \hat{h}_{\mathrm{BMA}}$ and $\hat{h}_{\mathrm{Select.}}$ respectively in the PB model, in the NL model, in the BMA and in the model selection framework. The predictive density for the latter is defined as

$$\hat{h}_{\mathrm{Select.}} = \mathbf{1}_{p_{\mathrm{PB}}(\mathscr{W})\geq 0.5}\hat{h}_{\mathrm{PB}} + \mathbf{1}_{p_{\mathrm{PB}}(\mathscr{W})<0.5}\hat{h}_{\mathrm{NL}}\,.$$

One of the main interest in multivariate extreme value theory may reside in the probability of an excess of a high threshold. In this study, we consider the probability of falling in the failure region $A(\mathbf{u})$ with $\mathbf{u} = (100, 100, 100)$. On the Fréchet scale, it corresponds to a marginal excess probability of roughly $\frac{1}{100}$. The quantity of interest Δ is thus a joint probability

$$\Delta(m, \underline{\theta}) = \mathbf{P}(\mathbf{X} > \mathbf{u}|m, \underline{\theta}) \simeq \nu(A(\mathbf{u})|m, \underline{\theta})$$

where $m \in \{\mathrm{PB}, \mathrm{NL}\}$ (see Section 2). The true probability is

$$\Delta(h_0) = \mathbf{P}_{h_0}(\mathbf{X} > \mathbf{u}) = \nu_0(A(\mathbf{u}))\,,$$

where ν_0 is the true exponent measure, which density in pseudo-polar coordinates is (*e.g.* on the region $\{r > 1\}$) $\mathrm{d}\nu_0 = d\frac{\mathrm{d}r}{r^2}h_0(\mathbf{w})\,\mathrm{d}\mathbf{w}$. Here, the approximation becomes an equality because the radii and angles are simulated independently from each other.

For the PB (*resp.* the true) density, $\nu(A(\mathbf{u})|\underline{\theta}, m)$ (*resp.* $\nu_0(A(\mathbf{u})$) is given by (4) and approximated by Monte-Carlo integration (since we can simulate angular samples from PB distributions and from the true one). In the NL model, it is simply given by (5).

The output of the Bayesian procedure in model m is, for a given data set $\mathscr{W}$, an approximate posterior sample $\{\underline{\theta}_m(t)\}_{1\leq t\leq T}$ [7]. This posterior is transformed into a posterior Δ-sample $\{\delta_m(t)\}_{1\leq t\leq T} = \{\Delta(\underline{\theta}_m(t))\}_{1\leq t\leq T} \in (0,1)$ of probabilities of failure, so that the posterior predictive distribution $\Delta^*(\pi_m|\mathscr{W})$ (see Section 4) on $(0,1)$ is approximated by the discrete cumulative distribution function (*cdf*)

$$\hat{F}_m(y) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\delta_m(t)\leq y} \quad (y \in (0,1))\,.$$

[7] Here, $T = 200$ after discarding the values from the burn-in period and thinning. The thinning interval is set to 100 to reduce the computational time

The posterior predictive *cdf* in the BMA is thus the weighted average

$$\hat{F}_{\mathrm{BMA}}(\,\cdot\,) = p_{\mathrm{PB}}(\mathscr{W})\,\hat{F}_{\mathrm{PB}}(\,\cdot\,) + p_{\mathrm{NL}}(\mathscr{W})\,\hat{F}_{\mathrm{NL}}(\,\cdot\,)\,.$$

We can now compare the different distributions *via* strictly proper scoring rules (Gneiting and Raftery, 2007), adapted to the case where the true distribution is known to be the Dirac mass at $\delta_0 = \Delta(h_0)$. Namely, we consider the continuous ranked probability score ($CRPS$), the predictive model choice criterion[8] ($PMCC$) and the interval score (IS_α) for the central $(1-\alpha) * 100\,\%$ interval based on predictive quantiles, with $\alpha = 0.1$. If $\hat{F}$, $\mathbf{E}_{\hat{F}}(\Delta)$, $\mathbf{Var}_{\hat{F}}(\Delta)$, $\hat{q}_{\alpha,l}$ and $\hat{q}_{\alpha,u}$ denote respectively the predictive *cdf* on $(0,1)$, the predictive mean and variance, and the predictive $\alpha/2$ and $1-\alpha/2$ quantiles, and if $\delta_0 \in (0,1)$ is the true value, these (negatively oriented) scores are

$$CRPS(\hat{F},\delta_0) = \int_{(0,1)} (\hat{F}(y) - \mathbf{1}_{\delta_0 \le y})^2 \,\mathrm{d}y\,, \tag{13}$$

$$PMCC(\hat{F},\delta_0) = (\mathbf{E}_{\hat{F}}(\Delta) - \delta_0)^2 + \mathbf{Var}_{\hat{F}}(\Delta)\,, \tag{14}$$

$$IS_\alpha(\hat{F},\delta_0) = \begin{cases} 2\alpha(\hat{q}_{\alpha,u} - \hat{q}_{\alpha,l}) + 4(\hat{q}_{\alpha,l} - \delta_0) & \text{if } \delta_0 \le \hat{q}_{\alpha,l}\,, \\ 2\alpha(\hat{q}_{\alpha,u} - \hat{q}_{\alpha,l}) & \text{if } \hat{q}_{\alpha,l} \le \delta_0 \le \hat{q}_{\alpha,u}\,, \\ 2\alpha(\hat{q}_{\alpha,u} - \hat{q}_{\alpha,l}) + 4(\delta_0 - \hat{q}_{\alpha,u}) & \text{if } \delta_0 > \hat{q}_{\alpha,u}\,. \end{cases} \tag{15}$$

Similarly to the logarithmic score, the 'best' model according to a given scoring rule is the one that minimizes the score.

Experimental setup.
Data sets are generated from Dirichlet mixture (DM) distributions (Boldi and Davison, 2007), which cover a wide variety of distributional shapes. The Dirichlet mixture parameters themselves are drawn according to a simulating rule described in the Appendix. Note that the hyper parameters for this simulating rule were chosen in order to grant significantly positive weights to both models.

We generate 20 DM parameters $\{\underline{\theta}_0^i\}_{1\le i\le 20}$ and for each $\underline{\theta}_0^i$, 5 data sets $\{\mathscr{W}_j^i\}_{1\le j\le 5}$ of size 80 each are generated according to the DM density h_0^i corresponding to $\underline{\theta}_0^i$.

For each of the 100 data sets $\{\mathscr{W}_j^i\}_{1\le i\le 20, 1\le j\le 5}$, separate inference is made in each framework 'fr' (here, fr represents the PB model, the NL model, the BMA and the model selection framework), yielding a density estimate $\hat{h}_j^i|_{fr}$, a *cdf* for the probability of failure $\hat{F}_j^i|_{fr}$ and posterior model weights which are approximated *via* (9). The posterior mode and the hessian matrix are approximated by numerical optimization.

Finally, the scores obtained by each framework are averaged over all the experiments (i,j).

Results.
The first panel of Table 1 shows the average scores (over the 100 data sets) obtained by each model, by the BMA and in the selection framework. The second panel

[8] This scoring rule is not proper in the general case but becomes so when the true distribution is a Dirac mass.

shows the average score differences[9] between the BMA and the three other possible approaches, together with an order of magnitude of the errors involved by the Monte-Carlo approximations used to compute the score differences between the BMA and the other approaches. More details about these errors are given in the Appendix.

For example, line ' BMA/ PB', column 'CRPS' corresponds to

$$CRPS(\text{BMA/PB}) = \sum_{i=1}^{20}\sum_{j=1}^{5} CRPS(\hat{F}_j^i|_{\text{BMA}}, \hat{\delta}^i) - CRPS(\hat{F}_j^i|_{\text{PB}}, \hat{\delta}^i),$$

where $\hat{\delta}^i$ is the Monte-Carlo estimate of the true probability of failure under the Dirichlet distribution with parameter $\underline{\theta}_0^i$, *i.e.*

$$\hat{\delta}^i = \frac{3}{N_{\text{mc}}} \sum_{N=1}^{N_{\text{mc}}} \min_{j\in 1:d} \left\{ \frac{w_{j,N}}{u_j} \right\} \; ; \quad \mathbf{w}_N \overset{i.i.d}{\sim} h_0^i \; ; \quad u_j = \frac{1}{100} \, .$$

Column 'LS' corresponds to

$$\sum_{i=1}^{20}\sum_{j=1}^{5} \hat{LS}(\hat{h}_j^i|_{\text{BMA}}) - \hat{LS}(\hat{h}_j^i|_{\text{PB}}),$$

where $\hat{LS}$ is given by (12).

TABLE 1. Comparison of mean scores with simulated data (error magnitude on score differences between parentheses).

Scores	*LS*	*CRPS*	*PMCC*	*IS*
PB	−107.48	24.04	20.97	45.06
NL	−106.07	21.34	**18.69**	38.08
BMA	**−108.39**	**21.33**	19.56	**37.21**
Select	−107.36	22.95	20.11	42.27
BMA/PB	−0.91 (0.32)	−2.7 (0.04)	−1.41 (0.05)	−7.85 (0.15)
BMA/NL	−2.33 (0.32)	−0.002 (0.09)	0.87 (0.08)	−0.87 (0.21)
BMA/Select	−1.03 (0.32)	−1.62 (0.02)	−0.55 (0.01)	−5.06 (0.12)

In terms of spectral density itself, the BMA approach obtains the best logarithmic score (first column, second panel). The logarithmic score obtained by the selection framework is also better than the ones obtained both by the PB and by the NL models. The gain is less obvious in terms of estimated probabilities of failure, probably because, for this kind of simulated data, the NL model obtains better average scores than the PB model (note that this tendency is reversed in terms of logarithmic scores). In any case, the BMA gives slightly, but consistently, better predictions, with respect to all the considered scores, than the model selection framework (line 7).

The disappointing aspect of these results is the fact that the relative gain or loss is low: roughly, between 1/100 and 1/10 depending on the considered score.

[9]The scores reported in each column have respectively been multiplied by $10^2, 10^5, 10^8$ and 10^5 to improve the readability of the numerical output.

4.3. **Example: Leeds data set.** We separately fit the PB and the NL model on the Leeds data set. Table 2 gathers results in terms of the transformed parameters in each model. $\hat{\theta}_{\text{post}}$ and $\hat{\sigma}_{\text{post}}$ denote the mean and standard deviation of the posterior sample issued from the Metropolis algorithm, $\hat{\theta}_{\text{mode}}$, $\hat{\sigma}_{\text{mode}}$, are the posterior mode and the 'standard deviation' represented by (with the notations of the Laplace approximation (9)) the squared root of the diagonal elements of the inverse hessian $\hat{\Sigma}$. The maximum likelihood estimates $\hat{\theta}_{mle}$ and the estimated standard errors $\hat{\sigma}_{mle}$ are also reported. Our Bayesian analysis corroborates the frequentist

TABLE 2. The PB and NL models fitted to Leeds data: Comparison between frequentist estimates and posterior summary statistics.

	PB model				NL model			
	$\log\beta_0$	$\log\beta_{12}$	$\log\beta_{13}$	$\log\beta_{23}$	$\text{logit}\,\alpha_0$	$\text{logit}\,\alpha_{12}$	$\text{logit}\,\alpha_{13}$	$\text{logit}\,\alpha_{23}$
$\hat{\theta}_{\text{post}}$	0.3	1.27	−0.35	1.22	0.22	0.89	4.57	1.19
$\hat{\sigma}_{\text{post}}$	0.14	0.43	0.23	0.42	0.09	0.49	1.69	0.64
$\hat{\theta}_{\text{mode}}$	0.3	1.3	−0.34	1.14	0.21	0.79	3.67	1.03
$\hat{\sigma}_{\text{mode}}$	0.14	0.43	0.22	0.42	0.09	0.45	1.23	0.42
$\hat{\theta}_{\text{mle}}$	0.3	1.32	−0.34	1.16	0.21	0.81	17.97	1.07
$\hat{\sigma}_{\text{mle}}$	0.14	0.43	0.22	0.43	0.09	0.47	2588.78	0.44

estimates. The unusually high standard deviation of the maximum likelihood estimate for $\text{logit}(\alpha_{13})$ is easily explained: the inverse logit link function is numerically flat (equal to 1) above 17, and $\text{logit}^{-1}(3.67) = 0.98$. As expected, adding a prior re-centres the estimates towards the origin, but the relative discrepancy between the Bayesian and frequentist modes (with respect to the standard deviation of the frequentist ones) is less than 0.12. Also, the posterior mode and mean are close to each other. This suggests that the asymptotic domain of validity of the Bernstein-Von Mises theorem (asymptotic normality of the posterior distribution, see *e.g.* van der Vaart, 2000) is approximately reached.

The posterior predictive spectral densities in the PB and NL models can be computed *via* (10). The squared dots in Figure 3 represent the data displayed in Figure 2. Each panel tells us the same main story, a lot of mass in the middle, more near the middle of the edges joining the pairs (PM10,NO) and (NO,NO2) than between the pair (PM10,NO2). This pattern roughly corresponds to the distribution of the observed angular points over the simplex. Still, the two panels have important differences. For example, the NL model assigns more mass to the regions near the vertices.

For this Leeds data set, the posterior weights are overwhelmingly in favour of the NL model. Table 3 gathers the posterior weights issued from the BIC approximation, the Laplace method and by simple Monte-Carlo integration (parameters are drawn from the prior).

For the Leeds data, BMA teaches us that a well-chosen four parameters NL model belonging to the large Nested Asymmetric Logistic family outperforms the PB model.

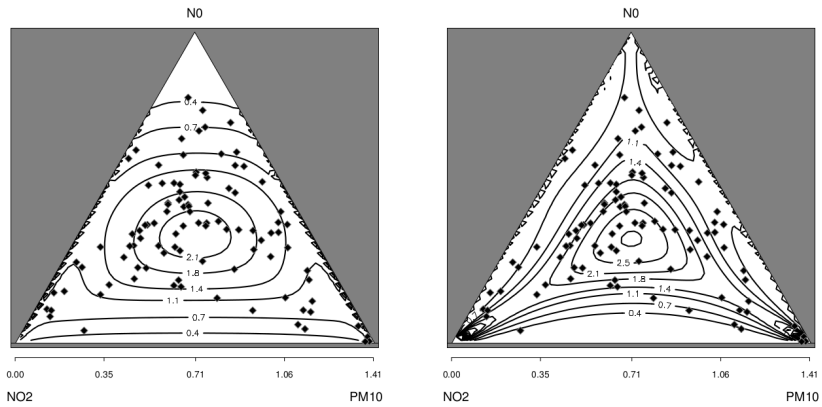

FIGURE 3. Leeds data: posterior predictive densities in the PB model (left panel) and the NL model (right panel).

TABLE 3. Leeds data set: posterior PB model weights and marginal likelihoods.
MC MC steps: 100×10^3.

	Laplace	BIC	Monte-Carlo
PB	$2.2\,10^{32}$	$4.1\,10^{32}$	$2.4\,10^{32}(\ 4.8\,10^{31})$
NL	$8.2\,10^{34}$	$1\,10^{35}$	$1.4\,10^{35}(\ 1.9\,10^{34})$
$\hat{p}_{\mathrm{PB}}(\mathbf{W})$	0.0027	0.004	0.0017

5. DISCUSSION

This article shows that it is feasible to implement a BMA approach for angular measure models. Simulation studies indicate that this approach can, at best, improve the predictive density estimates over each single model and at worst, be used as a selection tool by identifying a single one. For the four considered scoring rules, the gain represented by the BMA against the model selection framework is significant (in view of the Monte-Carlo errors) but moderate: the order of magnitude of the scores is unchanged.

For the sake of conciseness, we have only considered two models to be averaged. Future BMA roads would be to enlarge the dictionaries of parametric spectral families (e.g. for the PB model, Ballani and Schlather, 2011) and/or to extend the BMA framework to a mixture model setup, i.e. replacing the disjoint union parameter space by a product space. The resulting model would be more flexible, in the sense that the posterior mass would not have to concentrate on one single model for large sample sizes. As a drawback, the dimension of a product space is the sum of the dimensions of individual models, and the curse of dimensionality is likely to impose longer burn-in periods for MC MC algorithms. Also, one could not use posterior samples obtained in distinct models. We recall that, in this paper, we consider situations when separate inference has already been achieved, or can be made in a simple way, and where estimates are to be averaged. The main interest of the BMA approach is to offer a compromise between model flexibility

and parsimony: the estimated distribution (the posterior predictive) is a mixture, while inference is conducted in lower dimensional models.

Also, for our leading example, Heffernan and Tawn (2004)'s study suggests that the pairs ($S0_2$, NO) and (SO_2, PM10) might be asymptotically independent. It should thus be of interest to average general spectral measures associated with asymptotically independent models, as introduced by Ledford and Tawn (1996) and Ramos and Ledford (2009) in the bi-variate case, and extended to general multivariate problems in Qin *et. al.* (see footnote 4). The estimated distributions would not be max-stable anymore, but this would account for a potentially greater source of uncertainty than the one attached to model choice within the asymptotically dependent class.

Supplementary material

An R package is available at
`http://www.lsce.ipsl.fr/Phocea/Pisp/index.php?nom=anne.sabourin`

Acknowledgement

Part of this work has been supported by the EU-FP7 ACQWA Project (www.acqwa.ch), by the PEPER-GIS project, by the ANR-MOPERA project, by the ANR-McSim project and by the MIRACCLE-GICC project. The authors would like to thank Dan Cooley for his help with the PB model, and an anonymous referee for useful remarks.

Appendix 1: Bayesian PB and NL models

Simulation rule for the PB model.
Whereas Cooley *et al.* used an accept-reject method for simulation, the one that we propose here is direct. The PB density can be re-parametrized by setting $\rho_{ij} = w_i + w_j$, $w_{i/ij} = w_i/(w_i + w_j)$, $s_{ij} = w_{[-(i,j)]}/(1-\rho_{ij})$, where $w_{[-(i,j)]} = (w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{j-1}, w_{j+1}, \ldots, w_d)$.

The transformation $(\rho_{ij}, w_{i/ij}, s_{ij}) \mapsto w$ has Jacobian: $J = J(\rho_{ij}) = \rho_{ij}(1-\rho_{ij})^{d-3}$. Each beta function

$$h_{i,j}(\{w_{ij}, w_{i/ij}, s_{ij}\} \,|\beta_0, \beta_{ij}),$$

can be expressed within these new coordinates

$$h_{i,j}(\{\rho_{ij}, w_{i/ij}, s(w_{ij})\} \,|\beta_0, \beta_{ij}) \propto \rho_{ij}^{2\beta_0} (1-\rho_{ij})^{(d-2)\beta_0 - 1} \cdots \\ \cdots w_{i/ij}^{\beta_{ij}-1} \left(1 - w_{i/ij}\right)^{\beta_{i,j}-1} \frac{1}{J(\rho_{ij})},$$

which can be written with the standard R package notations as

$$\frac{1}{J(\rho_{ij})} \mathsf{dbeta}(\rho_{ij}, 2\beta_0+1, (d-2)\beta_0)\, \mathsf{dbeta}(w_{i/ij}, \beta_{i,j}, \beta_{i,j})\, \mathsf{ddirichlet}(s_{ij}, \mathsf{rep}(1, d-2)) .$$

The three factors correspond to two Beta distributions and one uniform distribution on the unit simplex of dimension $d-3$. The following algorithm produces the desired angular variables W according to the density $h_{\mathrm{PB}}(. \mid \beta_0, \beta_{ij})$.

Algorithm 1.

(1) *Choose uniformly a pair* $(i < j)$.

(2) *Generate independently the vectors* R_{ij}, $W_{i/ij}$ *and* S_{ij} *according to the Beta distributions* $\mathcal{B}e(2\beta_0+1,(d-2)\beta_0)$ *and* $\mathcal{B}e(\beta_{ij},\beta_{ij})$, *and the uniform Dirichlet distribution* $\mathcal{D}ir_{d-2}(1,\dots,1)$, *respectively.*

(3) *Define* W *as* $W_i = R_{ij}W_{i/ij}$, $W_j = R_{ij}(1-W_{i/ij})$ *and* $W_{[-(i,j)]} = (1-R_{ij})S_{ij}$.

Angular density in the NL model.
From Coles and Tawn (1991), Theorem 1, with our normalizing convention, the angular density on the simplex is $h_{\mathrm{NL}}(\mathbf{w}[\underline{\alpha}) = -\frac{1}{d}\,\partial_{1,2,3}V_{\mathrm{NL}}(\mathbf{x}|\underline{\alpha})|_{\mathbf{x}=\mathbf{w}}$, where we write $\partial_{i_1,\dots,i_k}(\,\cdot\,)$ the partial derivative with respect to $x_{i_1},\dots,x_{i_k}$.

Letting

(16)
$$U_{ij}(\mathbf{x}) = x_i^{\frac{-1}{\alpha_0\alpha_{ij}}} + x_i^{\frac{-1}{\alpha_0\alpha_{ij}}}\ (1\le i<j\le 3)\,, \qquad T(\mathbf{x}) = (U_{12}^{\alpha_{12}} + U_{13}^{\alpha_{13}} + U_{23}^{\alpha_{23}})\,(\mathbf{x})\,,$$

we have $V_{\mathrm{NL}} = 2^{-\alpha_0}T^{\alpha_0}(\mathbf{x})$, so that

$$\begin{aligned}\partial_{1,2,3}V_{\mathrm{NL}}(\mathbf{x}|\underline{\alpha}) = 2^{-\alpha_0}\alpha_0\,\big[&T^{\alpha_0-1}(\mathbf{x})\partial_{1,2,3}T(\mathbf{x}) + \cdots\\ &(\alpha_0-1)T^{\alpha_0-2}(\mathbf{x})\,\{\partial_1T(\mathbf{x})\partial_{2,3}T(\mathbf{x}) + \partial_2T(\mathbf{x})\partial_{1,3}T(\mathbf{x}) + \partial_3T(\mathbf{x})\partial_{1,2}T(\mathbf{x})\} + \cdots\\ &(\alpha_0-1)(\alpha_0-2)T^{\alpha_0-3}(\mathbf{x})\partial_1T(\mathbf{x})\partial_2T(\mathbf{x})\partial_3T(\mathbf{x})\big]\ .\end{aligned}$$

The simple and double partial derivatives are

$$\partial_iT(\mathbf{x}) = \frac{-1}{\alpha_0}\left(x_i^{\frac{-1}{\alpha_0\alpha_{ij}}-1}U_{ij}^{\alpha_{ij}-1} + x_i^{\frac{-1}{\alpha_0\alpha_{ik}}-1}U_{ik}^{\alpha_{ik}-1}\right) \tag{17}$$

and

$$\partial_{i,j}T(\mathbf{x}) = \frac{\alpha_{ij}-1}{\alpha_0^2\alpha_{ij}}\big(x_i\,x_j\big)^{\frac{-1}{\alpha_{ij}}-2}U_{ij}^{\alpha_{ij}-2}. \tag{18}$$

The third order derivative is thus zero. Finally, we have

$$\begin{aligned}h_{\mathrm{NL}}(\mathbf{w}|\underline{\alpha}) = \Bigg[&\frac{\alpha_0(1-\alpha_0)}{2^{\alpha_0}\,d}T^{\alpha_0-3}\cdots\\ &\cdots\Big\{\sum_{1\le i\ne j\ne k\le 3} T\partial_iT\partial_{j,k}T\ +\ (\alpha_0-2)\partial_1T\partial_2T\partial_3T\Big\}\Bigg]_{\mathbf{x}=\mathbf{w}}\end{aligned} \tag{19}$$

where all the terms are given in (16), (17) and (18).

Simulation method in the NL model.
We adapt here the method proposed by Stephenson (2003) to our context.

Algorithm 2.

(1) *Generate independently four positive alpha-stable variables* $S, S_{12}, S_{13}, S_{23}$, *with respective index* α_0, α_{12}, α_{13}, $\alpha_{23}\in(0,1)$, i.e. *with Laplace transform* $E(\exp(-tS)) = e^{-t^{\alpha_0}}$ *(*resp. $e^{-t^{\alpha_{ij}}}$*).*

(2) *For* $i\in\{1,2,3\}$:
   (a) *Simulate independently two standard exponentials* $E_{i,ij}, E_{i,ik}$.
   (b) *Set* $X_{i,ij} = \left[\left(\frac{S}{2}\right)^{\frac{1}{\alpha_{ij}}}\frac{S_{ij}}{E_{i,ij}}\right]^{\alpha_{ij}\alpha_0}$ *and* $X_{i,ik} = \left[\left(\frac{S}{2}\right)^{\frac{1}{\alpha_{ik}}}\frac{S_{ik}}{E_{i,ik}}\right]^{\alpha_{ik}\alpha_0}$.
   (c) *Set* $X_i = \max(X_{i,ij}, X_{i,ik})$.

Then, $\mathbf{X} = (X_1, X_2, X_3)$ *has unit Fréchet margins and a multivariate distribution belonging to the* NL *model (6).*

Proof. If $\mathbf{X}$ is generated according to the above algorithm, the conditional variables $\mathbf{X_{i,ij}}|\,(S = s, S_{ij} = s_{ij})$ are independent with distribution

$$\mathbf{P}\left(X_{i,ij} \leq x_i | s, s_{12}\right) = \exp\left(-s_{ij}\left(\frac{s}{2}\right)^{1/\alpha_{ij}}\left(\frac{1}{x_i}\right)^{1/(\alpha_0\alpha_{ij})}\right),$$

So that X has conditional distribution

$$\mathbf{P}\left(\mathbf{X} \leq \mathbf{x} | s, s_{12}\right) = \exp\left(-\sum_{1\leq i<j\leq 3} s_{ij}\left(\frac{s}{2}\right)^{1/\alpha_{ij}} \cdots \right.$$
$$\left. \cdots \left(\left(\frac{1}{x_i}\right)^{1/(\alpha_0\alpha_{ij})} + \left(\frac{1}{x_j}\right)^{1/(\alpha_0\alpha_{ij})}\right)\right).$$

Integrating with respect to the s_{ij}'s and s and using the Laplace transform property of positive α-stable variables yields the desired distribution function. □

The angular components $W_i = X_i/(X_1+X_2+X_3)$ follow immediately. By fixing a high threshold r_0 and retaining only the angular points corresponding to radii $R > r_0$, one obtains a sample on the simplex, approximately following angular distribution with density $h_{\mathrm{NL}}(.|\underline{\alpha})$.

Appendix 2: Results with simulated data from single models

Two data sets of 80 angular points each are simulated, one in the PB model, the other in the NL model. A 50 10^3-iteration Metropolis-Hastings is run, the last 20 10^3 values are kept.

The marginal posterior densities for the four parameters in the PB (*resp.* NL) model, obtained by a kernel smoothing of the *posterior* sample, are shown (solid lines) in Figure 4 (*resp.* Figure 5) , together with the prior densities (thin dotted lines) and the true parameters (vertical thick dotted lines). For all the parameters components, the posterior concentrates around the "true" value.

The posterior predictive density estimates are deduced from the posterior sample according to (10), and plotted in Figure 6. showing remarkable agreement between the estimated (solid lines), and the true distribution contours (dotted lines).

Basic summary statistics for the posterior samples are gathered in Table 4 (θ_0 stands for the "true" transformed parameter, see sub-section 4.3 for other notations), together with maximum likelihood estimates. The three approaches yield comparable results and the true parameter values lie at less than two standard deviations from their respective posterior mean estimates (except for the global dependence parameter α_0 in the NL model, where the discrepancy is about 2.4 for the three estimates).

Appendix 3: Simulation study

We give here a more complete account of the results obtained in Section 4.2.

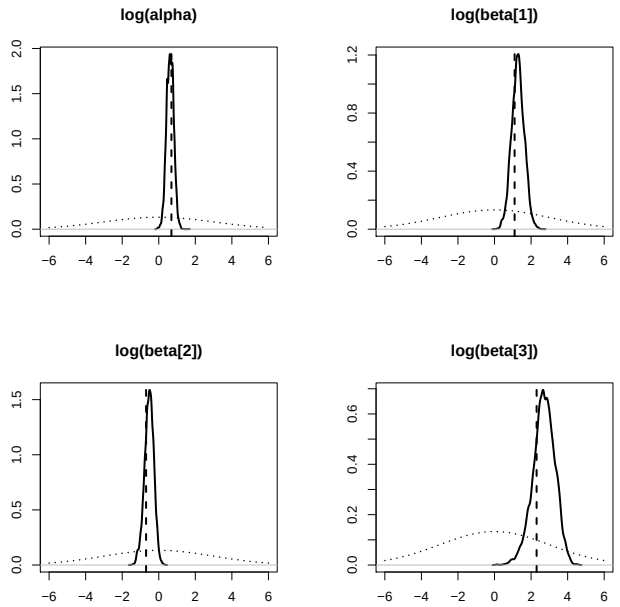

FIGURE 4. PB model: prior and posterior parameter marginal densities with simulated data. Upper left panel: $\text{logit}(\beta_0)$, upper right panel: $\text{logit}(\beta_{12})$, left and right lower panels: $\text{logit}(\beta_{13})$ and $\log(\beta_{23})$.

TABLE 4. The PB and NL models fitted to simulated data: Comparison between frequentist estimates and posterior summary statistics.

	PB model				NL model			
θ	$\log\beta_0$	$\log\beta_{12}$	$\log\beta_{13}$	$\log\beta_{23}$	$\text{logit}\,\alpha_0$	$\text{logit}\,\alpha_{12}$	$\text{logit}\,\alpha_{13}$	$\text{logit}\,\alpha_{23}$
θ_0	0.69	1.1	−0.69	2.3	0.41	−0.85	1.39	−0.41
$\hat{\theta}_{\text{post}}$	0.62	1.28	−0.53	2.7	0.15	−0.5	2.41	0.23
$\hat{\sigma}_{\text{post}}$	0.2	0.35	0.26	0.6	0.11	0.3	0.73	0.4
$\hat{\theta}_{\text{mode}}$	0.62	1.31	−0.52	2.77	0.14	−0.55	2.15	0.13
$\hat{\sigma}_{\text{mode}}$	0.19	0.34	0.25	0.58	0.11	0.28	0.61	0.35
$\hat{\theta}_{\text{mle}}$	0.62	1.32	−0.52	2.88	0.14	−0.56	2.25	0.13
$\hat{\sigma}_{\text{mle}}$	0.19	0.35	0.25	0.61	0.11	0.28	0.67	0.35

Dirichlet mixture model for spectral densities. Recall that the Dirichlet density, which we denote diri, can be parametrized by a mean vector $\boldsymbol{\mu} \in \mathbf{S}_d$ and

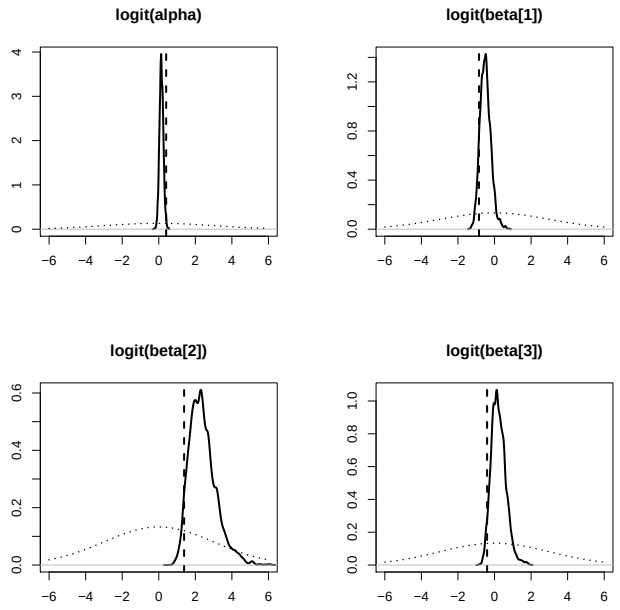

FIGURE 5. NL model: prior and posterior parameter marginal densities with simulated data. Upper left panel: $\log(\alpha_0)$, upper right panel: $\log(\alpha_{12})$, left and right lower panels: $\log(\alpha_{13})$ and $\log(\alpha_{23})$.

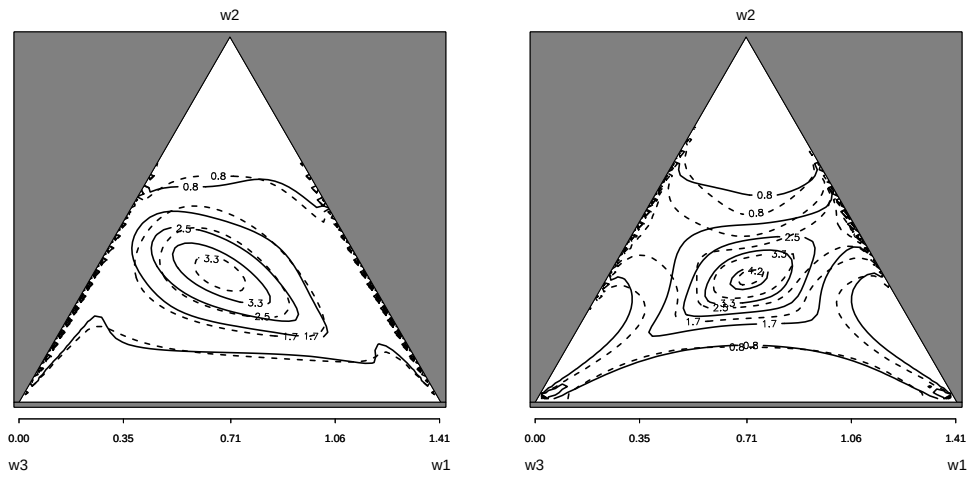

FIGURE 6. Angular measures: Simulation and estimation in the the PB model (left panel) and in the NL model(right panel).

a concentration parameter $\nu > 0$, so that

$$\forall \mathbf{w} \in \mathbf{S}_d,\ \mathrm{diri}(\mathbf{w} \mid \boldsymbol{\mu}, \nu) = \frac{\Gamma(\nu)}{\prod_{i=1}^{d} \Gamma(\nu\mu_i)} \prod_{i=1}^{d} w_i^{\nu\mu_i - 1}\,.$$

The Dirichlet mixture model is the family of finite mixtures of such densities, with positive weight vector $\boldsymbol{\omega} = \big(\omega_1, \ldots, \omega_K\big)$ $(K \geq 1)$ summing to one, concentration vector $\boldsymbol{\nu} = \big(\nu_1, \ldots, \nu_K\big)$ and mean matrix $\boldsymbol{\mu} = \big(\boldsymbol{\mu}_{\cdot,1}, \ldots, \boldsymbol{\mu}_{\cdot,K}\big)$ where $\boldsymbol{\mu}_{\cdot,k} = (\mu_{1,k}, \ldots, \mu_{d,k})$ is the mean vector for the kth mixture component. In this model, the mean constraint (2) is equivalent to

$$\forall i \in \{1, \ldots, d\},\ \sum_{k=1}^{K} \omega_k\, \mu_{i,k} = \frac{1}{d}\,. \tag{20}$$

Simulation of random Dirichlet mixture parameters and data sets. For our simulation, we generate 20 Dirichlet mixture parameters

$$\underline{\theta}_0^i = (\boldsymbol{\mu}^i_{\cdot,1:K}, \boldsymbol{\omega}^i_{1:K}, \boldsymbol{\nu}^i_{1:K})_{1 \leq i \leq 10}$$

with $K = 10$ components, so that (20) holds for all $\underline{\theta}_0^i$. Each $\underline{\theta}_0^i$ is generated as follows:

- For $k \in \{1, \ldots, K\}$, ν_k is generated under a truncated Gamma distribution with shape equal to 1.4 and scale equal to 10, with an upper bound set to 100.
- For $k \in \{1, \ldots, K-1\}$, $\boldsymbol{\mu}_{\cdot,k}$ is generated (independently) under a Dirichlet distribution with concentration parameter equal to 6 and a mean parameter set to $G_0 = (1/3, 1/3, 1/3)$, truncated to the region $\{\mathbf{w} \in \mathbf{S}_3 :\ \forall i \leq 3,\ w_i > \epsilon\}$ with $\epsilon = 1/100$.
- The first $K-1$ weights are constrained to be equal to each other and the location for the last kernel centre $\boldsymbol{\mu}_{\cdot,K}$ is set in in such a way that (20) holds while keeping the last weight ω_K as close to $1/K$ as possible.

For each $\underline{\theta}_0^i$, five data sets of size 80 each are generated under the corresponding Dirichlet mixture distribution. To avoid numerical errors, angular points with any coordinate less than 10^{-8} are rejected.

Error assessment for the mean differential scores. Since a lot of Monte-Carlo steps are involved in the differential score computations, the second part of Table 1 may only be interpreted as an order of magnitude for the errors. In the remainder of this subsection, an alternative *alt* denotes either the systematic choice of the PB or the NL model, or the model selection framework where the retained estimate is the one produce by the model with greatest posterior weight. If S is a scoring rule, $S(\mathrm{BMA}/alt)$ is the score difference between the BMAand the alternative.

Differential logarithmic score.
Here, we account for the error involved by the Monte-Carlo approximation (12).

For a given alternative *alt*, parameter $\underline{\theta}_0^i$ and data set $\mathbf{W}_j^i$, let $\hat{h}_j^i|_{\mathrm{BMA}}$ (*resp.* $\hat{h}_j^i|_{alt}$) the posterior predictive distributions in the BMA and in the alternative framework. Let $\hat{LS}(\hat{h}_j^i|_{\mathrm{BMA}})$ (*resp.* $\hat{LS}(\hat{h}_j^i|_{alt})$) be the Monte-Carlo estimate of the Logarithmic score as in (12), and let $\hat{\sigma}_j^i(\mathrm{BMA})$ (*resp.* $\hat{\sigma}_j^i(alt)$) be the classical Monte-Carlo error of the estimate. When i is fixed and j varies, the errors $\hat{\sigma}_j^i$ are not independent because they depend on the same Monte-Carlo sample. An estimated upper bound for the standard deviation of the differential score $\hat{LS}_j^i(\mathrm{BMA}/alt)$

is then $\hat{\sigma}_j^i = \hat{\sigma}_j^i(\mathrm{BMA}) + \hat{\sigma}_j^i(alt)$. This is conservative in the sense that this upper bound is only reached in the unrealistic case where $\hat{LS}_j^i(\mathrm{BMA})$ and $\hat{LS}_j^i(alt)$ have correlation equal to -1. In the same way, an upper bound for the standard deviation of the average (letting i fixed) is the average standard deviation: $\hat{\sigma}^i = \frac{1}{5}\sum_{j=1}^5 \hat{\sigma}_i^j$. Further, when i varies, the differential scores are independent from each other (*i.e.* if $i_1 \neq i_2$, $1 \leq j_1, j_2 \leq 5$, then $\hat{LS}_{j_1}^{i_1}(\mathrm{BMA}/alt)$ and $\hat{LS}_{j_2}^{i_2}(\mathrm{BMA}/alt)$ are independent). Consequently, an estimated upper bound for the variance of the average is $\hat{\sigma}(LS(\mathrm{BMA}/alt))^2 = \frac{1}{20^2}\sum_{i=1}^{20}\left[\hat{\sigma}^i\right]^2$. The errors reported in the first column, lines 9-11 of Table 1 are the squared roots of the latter quantity.

Failure region scores: $CRPS, PMCC$ and IS.
In this paragraph, the error concerns the approximation of the true probability of failure. Let $\hat{\delta}_0^i$, $\hat{\sigma}_0^i$ be respectively the mean Monte-Carlo estimate of the latter (see (4)), and its estimated standard deviation, for a given Dirichlet parameter $\underline{\theta}_0^i$. We define the boundaries of a typical centred error interval: $\delta_{\mathrm{inf}}^i = \hat{\delta}_0^i - \hat{\sigma}_0^i$, $\delta_{\mathrm{sup}}^i = \hat{\delta}_0^i + \hat{\sigma}_0^i$. Now, given a scoring rule S (one of the $CRPS, PMCC$ and IS rules) and an alternative alt, let $S^i(\mathrm{BMA}/alt, \delta_{\mathrm{inf}}^i)$ (*resp* $S^i(\mathrm{BMA}/alt, \delta_{\mathrm{sup}})$), be the mean differential score obtained between the BMA and framework alt, when the true failure probability is set to δ_{inf}^i (*resp.* δ_{sup}^i). For example, for the CRPS differential score between the PB model and the NL model, we set $CRPS^i(\mathrm{BMA/PB}, \delta_{\mathrm{sup}}^i) = \frac{1}{5}\sum_{j=1}^5 CRPS(\hat{F}_j^i|_{\mathrm{BMA}}, \delta_{\mathrm{sup}}^i) - CRPS(\hat{F}_j^i|_{\mathrm{PB}}, \delta_{\mathrm{sup}}^i)$.

An order of magnitude for the fluctuation of the partially averaged score $S^i(\mathrm{BMA}/alt, \hat{\delta}^i)$ is

$$err(S^i, alt) = \left|S^i(\mathrm{BMA}/alt, \delta_{\mathrm{inf}}^i) - S^i(\mathrm{BMA}/alt, \delta_{\mathrm{sup}}^i)\right|/2\,.$$

The final score $\hat{S}(\mathrm{BMA}/alt)$ is the average over $i \in \{1, \ldots, 20\}$ of the $S^i(\mathrm{BMA}/alt, \hat{\delta}_0^i)$'s, and the errors are independent when i varies. The heuristic error magnitude reported in the three last lines and last columns of Table 1 are thus (up to multiplication by the factor appearing in the column titles)

$$err(S, alt) = \left(\frac{1}{20}\sum_{i=1}^{20}\left[err(S^i, alt)\right]^2\right)^{1/2}.$$

References

Apputhurai, P. and Stephenson, A. (2011). Accounting for uncertainty in extremal dependence modeling using bayesian model averaging techniques. *Journal of Statistical Planning and Inference*, 141(5):1800–1807.

Ballani, F. and Schlather, M. (2011). A construction principle for multivariate extreme value distributions. *Biometrika*, 98(3).

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004). *Statistics of extremes: Theory and applications*. John Wiley & Sons: New York.

Berk, R. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58.

Boldi, M.-O. and Davison, A. C. (2007). A mixture model for multivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):217–229.

Coles, S. and Tawn, J. (1991). Modeling extreme multivariate events. *JR Statist. Soc. B*, 53:377–392.

Cooley, D., Davis, R., and Naveau, P. (2010). The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis*, 101(9):2103–2117.

Cowles, M. and Carlin, B. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, pages 883–904.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory, An Introduction*. Springer Series in Operations Research and Financial Engineering.

Einmahl, J., de Haan, L., and Piterbarg, V. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *The Annals of Statistics*, 29(5):1401–1423.

Einmahl, J. and Segers, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics*, 37(5B):2953–2989.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *IN BAYESIAN STATISTICS*, pages 169–193. University Press.

Gneiting, T. and Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gudendorf, G. and Segers, J. (2011). Nonparametric estimation of an extreme-value copula in arbitrary dimensions",. *Journal of Multivariate Analysis*, 102:37–47.

Guillotte, S., Perron, F., and Segers, J. (2011). Non-parametric bayesian inference on bivariate extremes. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 73:377–406.

Gumbel, E. (1960). Distributions des valeurs extrêmes en plusieurs dimensions. *Publ. Inst. Statist. Univ. Paris*, 9:171–173.

Heffernan, J. and Tawn, J. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

Heidelberger, P. and Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Commun. ACM*, 24:233–245.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical science*, 14(4):382–401.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the american statistical association*, pages 773–795.

Kass, R., Tierney, L., and Kadane, J. (1990). The validity of posterior expansions based on laplace's method. *Bayesian and Likelihood methods in Statistics and Econometrics*, 7:473–488.

Kleijn, B. and van der Vaart, A. (2006). Misspecification in infinite-dimensional bayesian statistics. *The Annals of Statistics*, 34(2):837–877.

Ledford, A. and Tawn, J. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Madigan, D. and Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

Raftery, A., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174.

Ramos, A. and Ledford, A. (2009). A new class of models for bivariate joint tails. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):219–241.

Resnick, S. (1987). *Extreme values, regular variation, and point processes, volume 4 of Applied Probability. A Series of the Applied Probability Trust.* Springer-Verlag, New York.

Resnick, S. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling.* Springer Series in Operations Research and Financial Engineering.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation.* Springer Verlag.

Stephenson, A. (2003). Simulating multivariate extreme value distributions of logistic type. *Extremes*, 6(1):49–59.

Tawn, J. (1990). Modelling multivariate extreme value distributions. *Biometrika*, 77(2):245.

van der Vaart, A. (2000). *Asymptotic statistics (Cambridge series in statistical and probabilistic mathematics).* Cambridge University Press.

(Anne Sabourin) Laboratoire des Sciences du Climat et de l'Environnement, CNRS-CEA-UVSQ, 91191 Gif-sur-Yvette, France or
Universite de Lyon, CNRS UMR 5208,Universite de Lyon 1, Institut Camille Jordan ,43 blvd. du 11 novembre 1918, F-69622 Villeurbanne cedex , France
E-mail address: anne.sabourin@lsce.ipsl.fr

(Philippe Naveau) CNRS-CEA-UVSQ, 91191 Gif-sur-Yvette, France
E-mail address: philippe.naveau@lsce.ipsl.fr

(Anne-Laure FOUGÈRES) Universite de Lyon,CNRS UMR 5208, Universite de Lyon 1, Institut Camille Jordan,43 blvd. du 11 novembre 1918; F-69622 Villeurbanne cedex, France
E-mail address: fougeres@math.univ-lyon1.fr

Chapitre 3

Un modèle de mélange de Dirichlet pour les extrêmes

Se limiter au cadre paramétrique, aussi élargi soit-il par la combinaison finie de modèles, revient à faire une hypothèse forte sur la structure de dépendance des extrêmes. Sans revenir sur les motivations déjà présentées en introduction et reprises dans l'article qui suit, disposer d'un cadre bayésien non-paramétrique pour modéliser les excès au-dessus d'un seuil élevé paraît souhaitable, alors que la littérature est rare sur le sujet. Boldi et Davison (2007) ouvrent la voie de la modélisation de la mesure angulaire sur le simplexe par un mélange de Dirichlet à nombre arbitraire de composants. L'inférence bayésienne dans ce modèle est en théorie possible, mais délicate en pratique, de par la présence de la contrainte de moments (1.16). A vu de l'instabilité des résultats obtenus en dimension plus grande que trois, sur des données simulées, on pouvait raisonnablement se demander si les conditions pour la consistance de la loi a posteriori, définie en section 1.3.3, étaient réunies. Il n'était pas non plus évident que l'algorithme d'échantillonnage avait les bonnes propriétés d'ergodicité, au sens de la section 1.3.4. Ce chapitre est consacré à la re-paramétrisation de ce modèle de mélange, de sorte que la contrainte soit automatiquement satisfaite, à l'établissement des propriétés de consistance de la loi a posteriori et à l'implémentation d'un algorithme à sauts réversibles permettant d'échantillonner cette dernière en un temps acceptable lorsque la dimension de l'espace des observations est modérée. L'ensemble de ces résultats a été accepté pour publication en 2013 dans la revue *Computational Statistics and Data Analysis*.

Bayesian Dirichlet mixture model for multivariate extremes: a re-parametrization.

A. SABOURIN[a,b], P. NAVEAU[a]

[a] *Laboratoire des Sciences du Climat et de l'Environnement, CNRS-CEA-UVSQ, 91191 Gif-sur-Yvette, France*

[b] *Université de Lyon, CNRS UMR 5208, Université de Lyon 1, Institut Camille Jordan, 43 blvd. du 11 novembre 1918, F-69622 Villeurbanne cedex , France*

Abstract

The probabilistic framework of extreme value theory is well-known: the dependence structure of large events is characterized by an angular measure on the positive orthant of the unit sphere. The family of these angular measures is non-parametric by nature. Nonetheless, any angular measure may be approached arbitrarily well by a mixture of Dirichlet distributions. The semi-parametric Dirichlet mixture model for angular measures is theoretically valid in arbitrary dimension, but the original parametrization is subject to a moment constraint making Bayesian inference very challenging in dimension greater than three. A new unconstrained parametrization is proposed. This allows for a natural prior specification as well as a simple implementation of a reversible-jump MCMC. Posterior consistency and ergodicity of the Markov chain are verified and the algorithm is tested up to dimension five. In this non identifiable setting, convergence monitoring is performed by integrating the sampled angular densities against Dirichlet test functions.

Keywords: multivariate extremes, semi parametric Bayesian inference, mixture models, reversible-jump algorithm

1. Introduction

Estimating the dependence among extreme events in a multivariate context has proven to be of great importance for risk management policies. The main

Email addresses: `anne.sabourin@lsce.ipsl.fr` (A. SABOURIN), `philippe.naveau@lsce.ipsl.fr` (P. NAVEAU)

probabilistic framework of multidimensional extreme value theory is well-known, but inference and model choice remain an active research field. The dependence structure of multivariate extreme value distributions is characterized by the so-called *spectral measure* (or *angular measure*), which is defined on the unit positive quadrant of the observations space. The non-parametric nature of this angular measure calls for fully non-parametric methods. Still, a moment constraint has to be satisfied and this restriction makes modeling and inference complex.

In a frequentist context, an empirical spectral measure estimator has been proposed by Einmahl et al. (2001) and amended by Einmahl and Segers (2009), for the two dimensional case. Weak convergence of a rescaled version of the empirical measure has been established, but the intricate form of the limit law does not provide, to our understanding, a simple way to derive asymptotic confidence bounds. In a similar context, de Carvalho et al. (2013) provide a simpler Euclidean likelihood estimator but an explicit expression for the asymptotic variance is still missing. The recurrence of such difficulties in the field of multivariate extremes is a strong argument in favor of Bayesian methods. To your knowledge, Guillotte et al. (2011) are the only authors having implemented a fully non-parametric Bayesian model, and the latter is only applicable to the bi-variate case.

Boldi and Davison (2007) were the first ones to adapt the elegant Dirichlet mixture (DM) framework to multivariate extreme values and to provide posterior predictive distributions in this context. This semi-parametric model (with varying number of mixture components) is designed for any sample space's dimension and weakly dense in the set of admissible angular measures. As posteriors were very difficult to sample from, Boldi and Davison (2007) also resorted to maximum-likelihood methods based on an EM algorithm and they concluded that "one practical drawback with the approach stems from the use of simulation algorithms, which may converge slowly unless they have been tuned. A second is that the number of parameters increases rapidly with the number of mixture components, so model complexity must be sharply penalized through an information criterion or a prior on the number of mixture components". One other key point about this past work is that Bayesian estimation in dimension greater than three was rendered very delicate by the low convergence rate of the reversible-jump Metropolis algorithm used to approximate the posterior distribution. Most of the difficulties they encountered were linked to the above mentioned moment constraint. Still, a workable spectral estimator based on Dirichlet distributions will be a valuable semi-parametric tool for Bayesian practitioners who would like to analyze multivariate extremes of moderate dimensions (i.e. around five).

Following Boldi and Davison's steps, we propose in this paper a novel parametrization of the DM model. One strong advantage of this parametrization resides in

the fact that the moment constraint is automatically satisfied. This allows to construct a prior in a relatively simple way (section 3), and it is verified that the posterior is consistent for a large class of 'true' distributions. A trans-dimensional *Metropolis-within-Gibbs* algorithm is implemented (section 4) to approach the posterior distribution. In practice, assuming that the maximum number of clusters within the mixture is below 15 (a reasonable hypothesis for most applications), it becomes possible to make accurate Bayesian inferences for at least five dimensional data sets (see section 7).

Theoretical ergodicity properties of the algorithm are investigated in section 5 and section 6 deals with the important issue of empirical convergence assessment. Like in any other mixture model, the parameters are not identifiable, and the monitored quantity cannot be a parameter component. Instead, convergence of the *densities* can be checked, and we propose an approach based on the use of well chosen Dirichlet test functions to be integrated against the Dirichlet mixture densities generated by the algorithm. In addition, this method allows goodness-of-fit checking. In section 7, a simulation study is performed with three- and five- dimensional data sets, in order to compare our algorithm with Boldi and Davison's one, in terms of mixing properties and predictive accuracy. We also fit the Dirichlet mixture model to air quality measurements, recorded in the city of Leeds, UK, during the winter season, years 1994-1998. This data set is available at `http://www.airquality.co.uk` and has already been studied by Cooley et al. (2010), Heffernan and Tawn (2004), Boldi and Davison (2007) and Sabourin et al. (2013) and we comment our results with respect to Boldi and Davison (2007)'s approach. Another simulation study is performed to assess the impact of the prior specification. Finally, comparison is made with Guillotte et al. (2011)'s non-parametric Bayesian model in a bi-variate setting. Our results are discussed in section 8.

2. Background and notations

2.1. Multivariate extremes and spectral measure

Multivariate extreme value theory aims at characterizing the joint behavior of extreme events such as block maxima or multivariate excesses above a threshold (Beirlant et al., 2004; de Haan and Ferreira, 2006; Resnick, 1987, 2007). Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a positive random vector of size d. If the uni-variate marginal distributions are known, there is no loss of generality in assuming each of them to be unit-Fréchet distributed $P(X_i \leq x) = \exp\left(-\frac{1}{x}\right)$, for $i = 1, \ldots, d$. Concerning the multivariate dependence description, it is convenient to introduce the L^1 norm $R = X_1 + \cdots + X_d$ and to represent $\mathbf{X}$ in polar coordinates, letting R be the radial

component and $\mathbf{W} = \mathbf{X}/R$ the angular one. Thus, $\mathbf{W}$ corresponds to a random point on the $d-1$ dimensional unit simplex $\mathbf{S}_d = \big\{\mathbf{w} = (w_1, \cdots, w_d) \ : w_i \geq 0 \ w_1 + \cdots + w_d = 1\big\}$.

A major result of multivariate extreme value theory is that, under mild assumptions (see *e.g.* Resnick, 1987, multivariate regular variation), the radial and angular components become independent for large R's. More precisely, with our choice of unit Fréchet margins, the condition is that the cumulative distribution function (*cdf*) of $\mathbf{X}$ be in the domain of attraction of a max-stable distribution G, *i.e.* that there exist a non degenerate *cdf* G such that the limit $P^t(\mathbf{X} \leq t\mathbf{x})$ goes to $G(\mathbf{x})$, as $t \to \infty$. This implies $G^t(t\mathbf{x}) = G(\mathbf{x})$ for all $t > 0$. In such a case, there is a *spectral probability measure* H defined on $\mathbf{S}_d$, such that for any Borelian subset B of $\mathbf{S}_d$, $P(\mathbf{W} \in B, R > r) \underset{r\to\infty}{\sim} r^{-1}H(B)$, so that

$$P(\mathbf{W} \in B \mid R > r) \underset{r\to\infty}{\longrightarrow} H(B). \tag{1}$$

Thus, H represents the distribution of the angular components for asymptotically large R's. This measure has to satisfy the moment constraint

$$\text{for all } i = 1, \ldots, d, \ \int_{\mathbf{S}_d} w_i \, \mathrm{d}H(\mathbf{w}) = \frac{1}{d}. \tag{2}$$

Conversely, any probability measure H satisfying (2) is a valid spectral measure for a multivariate extreme value distribution G. In other words, H is a valid spectral measure if and only if its center of mass lies at the centroid of the unit simplex. In this paper, we focus on angular measures which mass is concentrated on the interior of the unit simplex, denoted $\overset{\circ}{\mathbf{S}}_d$, and which admit densities with respect to the Lebesgue measure $\mathrm{d}w_1 \cdots \mathrm{d}w_{d-1}$ on the Euclidean plane of dimension $d-1$. The simplex is parametrized by $\{(w_1, \ldots, w_{d-1}) \ : \ w_i \geq 0 \, ; \, \sum_{i=1}^{d-1} w_i \leq 1\}$.

2.2. Dirichlet mixture model (Boldi and Davison, 2007)

Besides condition (2), there is no other constraint on H. In terms of modeling, this strongly favors non-parametric, or semi-parametric models. As H lives on the interior of the unit-simplex, the Dirichlet mixtures family appears as the ideal candidate. We recall that a Dirichlet density, which we denote diri, can be parametrized by a mean vector $\boldsymbol{\mu} \in \overset{\circ}{\mathbf{S}}_d$ and a concentration parameter $\nu > 0$, so that

$$\forall \mathbf{w} \in \mathbf{S}_d, \ \mathrm{diri}(\mathbf{w} \mid \boldsymbol{\mu}, \nu) = \frac{\Gamma(\nu)}{\prod_{i=1}^d \Gamma(\nu\mu_i)} \prod_{i=1}^d w_i^{\nu\mu_i - 1}.$$

A k-component Dirichlet mixture density is a finite mixture

$$h_{(\boldsymbol{\mu},\mathbf{p},\boldsymbol{\nu})}(\mathbf{w}) = \sum_{m=1}^{k} p_m \,\mathrm{diri}(\mathbf{w} \mid \boldsymbol{\mu}_{\cdot,m}, \nu_m),$$

with positive weight vector $\mathbf{p} = \big(p_1, \ldots, p_k\big)$ summing to one, concentration vector $\boldsymbol{\nu} = \big(\nu_1, \ldots, \nu_k\big)$ and mean matrix $\boldsymbol{\mu} = \big(\boldsymbol{\mu}_{\cdot,1}, \ldots, \boldsymbol{\mu}_{\cdot,k}\big)$, where $\boldsymbol{\mu}_{\cdot,m} = (\mu_{1,m}, \ldots, \mu_{d,m})$ is the mean vector for the mth mixture component. The moment constraint (2) is equivalent to

$$\sum_{m=1}^{k} p_m \,\mu_{i,m} = \frac{1}{d}, \text{for all } i = 1, \ldots, d. \tag{3}$$

This leads to the Ψ-parametrization proposed and studied by Boldi and Davison (2007) as a disjoint union:

$$\Psi = \coprod_{k \geq 1} \Psi_k\,, \text{ with } \Psi_k = \big\{\boldsymbol{\psi} = \big(\boldsymbol{\mu}_{\cdot,1:k}, p_{1:k}, \nu_{1:k}\big) : \text{ (3) holds}\big\}\ .$$

Here, the vector $\boldsymbol{\mu}_{\cdot,q:r}$ denotes $\big(\boldsymbol{\mu}_{\cdot,q}, \ldots, \boldsymbol{\mu}_{\cdot,r}\big)$ for $q \leq r$. This type of notation will be used throughout this work, e.g. $p_{q:r}$ means $(p_q, \ldots, p_r)$. Unless otherwise mentioned, $\|\cdot\|$ denotes the Euclidean norm on $\mathbf{R}^d$ while $\|\cdot\|_1$ stands for the L^1 norm.

The weak density of such mixtures within the space of admissible angular measures, proved by Boldi and Davison (2007), renders this model very attractive in terms of flexibility. However, in a Bayesian context, specifying an adequate prior distribution for $\boldsymbol{\mu} = \boldsymbol{\mu}_{\cdot,1:k}$ and $\mathbf{p} = p_{1:k}$ subject to (3) is challenging. Boldi and Davison (2007) conditioned $\boldsymbol{\mu}$ upon $\mathbf{p}$. The prior on $\boldsymbol{\mu}$ was then defined component by component, on the open set

$$\left\{\mu_{1:d-1,1:k-1} : \forall 1 \leq m < k, \sum_{i=1}^{d-1} \mu_{i,m} < 1 \text{ and } \forall 1 \leq i < d, \sum_{m=1}^{k-1} p_m \mu_{i,m} < \frac{1}{d}\right\},$$

by successive conditioning, each component being uniformly distributed on the largest interval keeping (3) satisfied. Besides a minor error on the admissible bounds of such an interval (see Appendix G.1 for details), doing so introduced some asymmetry in $\boldsymbol{\mu}$'s *prior* distribution: in particular, the coordinates $\mu_{i,m}$ ($i = 1, \ldots, d$) of a given mean vector $\boldsymbol{\mu}_{\cdot,m}$ were not exchangeable in their model and

the prior was concentrated in a relatively small region of the space of admissible mixtures. This might partly explain the low convergence rate of their reversible jump algorithm: such an asymmetric concentration may lead to the rejection of many proposals and to a low acceptance rate. Below, we address this issue by proposing an alternative parametrization such that constraint (2) is automatically satisfied. This allows a natural prior specification in which space coordinates play symmetrical roles.

3. Unconstrained Dirichlet mixture model

3.1. Re-parametrization

Our goal is to replace the weight vector $\mathbf{p}$ and the "last" mean vector $\boldsymbol{\mu}_{\cdot,k}$ by eccentricities $\mathbf{e} = (e_1, \ldots, e_{k-1})$, between zero and one. Those e_m's are sequentially defined and indicate departure from centrality induced by decreasing subsets of mixture components. Thus, (3) is automatically satisfied and the parameter space for k-mixtures is a "rectangular" subset of $\mathbf{S}_d^{k-1} \times (0,1)^{k-1} \times (\mathbf{R}^+)^k$.

Let us go into details: suppose one wants to construct a k-components DM density $h_{(\boldsymbol{\mu},\mathbf{p},\boldsymbol{\nu})}$ satisfying (3). For $m \in \{0, \ldots, k-1\}$, we introduce, as an intermediate variable, the center of mass $\boldsymbol{\gamma}_m$ of the $k-m+1$ last components

$$\boldsymbol{\gamma}_m = \rho_m^{-1} \sum_{j=m+1}^{k} p_j \, \boldsymbol{\mu}_{\cdot,j},$$

where $\rho_m = \sum_{j=m+1}^{k} p_j = 1 - \sum_{j=1}^{m} p_j$ $(m \geq 1)$, and $\rho_0 = 1$. From (3), we know that $\boldsymbol{\gamma}_0 = (1/d, \ldots, 1/d)$, and, by associativity of the center of mass, $\boldsymbol{\gamma}_m$ may also be expressed in terms of the *preceding* mixture components:

$$\boldsymbol{\gamma}_m = \rho_m^{-1} \left(\boldsymbol{\gamma}_0 - \sum_{j=1}^{m} p_j \, \boldsymbol{\mu}_{\cdot,j} \right) . \tag{4}$$

By associativity again, for $m = 1$, we have

$$\begin{aligned} \boldsymbol{\gamma}_0 &= p_1 \, \boldsymbol{\mu}_{\cdot,1} + \sum_{j=2}^{k} p_j \, \boldsymbol{\mu}_{\cdot,j} \\ &= p_1 \, \boldsymbol{\mu}_{\cdot,1} + \rho_1 \, \boldsymbol{\gamma}_1 \, . \end{aligned}$$

Visually, this means that $\boldsymbol{\gamma}_0$ is located on the line segment joining $\boldsymbol{\gamma}_1$ and $\boldsymbol{\mu}_{\cdot,1}$ (see Figure 1, with $m = 1$, on the two-dimensional simplex $\mathbf{S}_3$), *i.e.* that $\boldsymbol{\gamma}_1$ lies on

the half line $\mathcal{D}_1 = [\boldsymbol{\gamma}_0, \boldsymbol{\mu}_{\cdot,1})$, inside the simplex. If I_1 is the intersection between $\mathcal{D}_1$ and the boundary of the simplex, it is clear that one can use a number e_1 between 0 and 1 to determine the position of $\boldsymbol{\gamma}_1$ on the segment $[\boldsymbol{\gamma}_0, I_1]$. Namely, set

$$e_1 = \frac{\|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_0\|}{\|I_1 - \boldsymbol{\gamma}_0\|}\,.$$

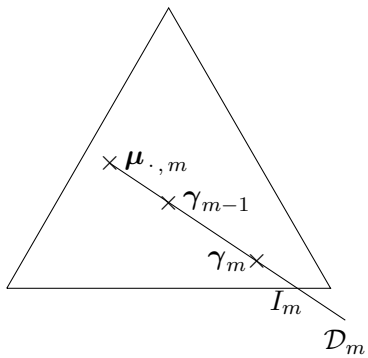

Figure 1: Sequential construction of the partial centers of mass on the two-dimensional simplex $\mathbf{S}_3$ at step m. The simplex points γ_m, γ_{m-1} and m^{th} mean vector $\boldsymbol{\mu}_{\cdot,m}$, as defined in Proposition 1, belong to a common line $\mathcal{D}_m$ and γ_{m-1} lies between γ_m and $\boldsymbol{\mu}_{\cdot,m}$, so that (7) holds for some eccentricity parameter $e_m \in (0,1)$.

At this stage, given $\boldsymbol{\mu}_{\cdot,1}$ and e_1, one can deduce the location of $\boldsymbol{\gamma}_1$ and elementary algebra provides relative weights p_1 and ρ_1 respectively assigned to $\mu_{\cdot,1}$ and $\boldsymbol{\gamma}_1$. In a second step, as above, we have

$$\boldsymbol{\gamma}_1 = p_2\,\boldsymbol{\mu}_{\cdot,1} + \rho_2\,\boldsymbol{\gamma}_2\,.$$

The argument can be repeated to obtain recursively the subsequent centers of mass $\boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_{k-1}$ and weights $p_2, \ldots, p_{k-1}$, $\rho_2, \ldots, \rho_{k-1}$, given $k-1$ Dirichlet mean vectors $\boldsymbol{\mu}_{\cdot,1:k-1}$ and eccentricities $e_{1:k-1}$, *via*

$$\begin{cases} \boldsymbol{\gamma}_m = \boldsymbol{\gamma}_{m-1} + e_m(I_m - \boldsymbol{\gamma}_m) \\ p_m = \rho_{m-1}\frac{\|\boldsymbol{\gamma}_m - \boldsymbol{\gamma}_{m-1}\|}{\|\boldsymbol{\gamma}_m - \mu_{\cdot,m}\|} \\ \rho_m = \rho_{m-1} - p_m \end{cases}$$

Finally, from the definition, $\boldsymbol{\gamma}_{k-1} = \boldsymbol{\mu}_{\cdot,k}$ and $p_k = \rho_{m-1}$.

Roughly speaking, e_m rules the eccentricity induced by $\boldsymbol{\mu}_{\cdot,m}$ onto the subsequent partial center of mass $\boldsymbol{\gamma}_m$, relatively to the current one $\boldsymbol{\gamma}_{m-1}$. It also determines the weight to be attributed to $\boldsymbol{\mu}_{\cdot,m}$: for a small e_m, $\boldsymbol{\gamma}_{m-1}$ and $\boldsymbol{\gamma}_m$ are close to each other, *i.e.* the departure from $\boldsymbol{\gamma}_{m-1}$ induced by $\boldsymbol{\mu}_{\cdot,m}$ is small, so that p_m is also small.

It must be noted that the parametrization is valid only if

$$\boldsymbol{\gamma}_{m-1} \neq \boldsymbol{\mu}_{\cdot,m}, \text{ for all } m \in \{1,\ldots,k-1\}. \tag{5}$$

This condition is satisfied for all $\boldsymbol{\mu}_{\cdot,1},\ldots,\boldsymbol{\mu}_{\cdot,k-1}$ out of a nowhere dense subset of $\mathbf{S}_d^{k-1}$. In practice, it will be almost surely satisfied if one chooses any absolutely continuous prior for the $\boldsymbol{\mu}_{\cdot,m}$'s.

For computational purposes, analytical expressions for the $\boldsymbol{\gamma}_m$'s are needed in order to derive the weights and the last mean vector $\boldsymbol{\mu}_{\cdot,k}$. We thus introduce the positive scalar

$$T_m = \sup\left\{t \geq 0 : \ \boldsymbol{\gamma}_{m-1} + t\left(\boldsymbol{\gamma}_{m-1} - \boldsymbol{\mu}_{\cdot,m}\right) \in \mathbf{S}_d\right\} \left(m \in \{1,\ldots,k-1\}\right), \tag{6}$$

so that $I_m = \boldsymbol{\gamma}_{m-1} + T_m(\boldsymbol{\gamma}_{m-1} - \boldsymbol{\mu}_{\cdot,m})$, and that

$$\boldsymbol{\gamma}_m = \boldsymbol{\gamma}_{m-1} + e_m T_m\left(\boldsymbol{\gamma}_{m-1} - \boldsymbol{\mu}_{\cdot,m}\right). \tag{7}$$

It is shown in Appendix A, that

$$T_m = \min_{i \in \mathcal{C}_m} \frac{\gamma_{i,m-1}}{\mu_{i,m} - \gamma_{i,m-1}}, \tag{8}$$

where $\mathcal{C}_m$ is the index set $\left\{i \in \{1,\ldots,d\} \ : \ \gamma_{i,m-1} - \mu_{i,m} < 0\right\}$. The following proposition summarizes the argument.

Proposition 1. *Let $h_{(\boldsymbol{\mu},\mathbf{p},\boldsymbol{\nu})}$ be a k-component* DM *density satisfying* (3) *and* (5), *with partial centers of mass $\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_{k-1}$ defined in* (4). *Let $\{T_m \ : \ 1 \leq m \leq k-1\}$ be the positive scalars introduced in* (6).

Then, we have $\boldsymbol{\gamma}_0 = (1/d,\ldots,1/d)$, each T_m is given by (8), *and there exists* $k-1$ eccentricity parameters $\left(e_1,\ldots,e_{k-1}\right) \in (0,1)^{k-1}$ *such that* (7) *holds for all* $m \in \{1,\ldots,k-1\}$.

Conversely, suppose that $\boldsymbol{\mu}_{\cdot,1:k-1} \in (\overset{\circ}{\mathbf{S}}_d)^{k-1}$ and $e_{1:k-1} \in (0,1)^{k-1}$ satisfying (5) *are given, together with a concentration vector $\nu_{1:k}$, $\nu_i > 0$.*

Then, one may successively define centers of mass $\{\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_{k-1}\}$ through (7), *where T_m is given by* (8); *together with weights $p_{1:k-1}, \rho_{0:k-1}$* via $\rho_0 = 1$ *and for*

$1 \leq m \leq k-1$,

$$p_m = \rho_{m-1} \frac{e_m T_m}{1 + e_m T_m} \; ; \quad \rho_m = \rho_{m-1} - p_m \, . \tag{9}$$

Setting the last mean vector and weight to $\boldsymbol{\mu}_{\cdot,k} = \boldsymbol{\gamma}_{k-1}$ *and* $p_k = \rho_{k-1}$ *, the* DM *parameters* $(\boldsymbol{\mu}, \mathbf{p}, \boldsymbol{\nu})$ *satisfy the moment constraint* (3) *and the* DM *density* $h_{\boldsymbol{\mu},\mathbf{p},\boldsymbol{\nu}}$ *is an admissible angular measure.*

In other words,the re-parametrized model is in one-to one correspondence with the original DM model introduced by Boldi and Davison (2007) (if the latter is restricted to the non degenerate mixtures verifying (5)). The *unconstrained parameter space* for the DM model can now be defined as a disjoint union

$$\boldsymbol{\Theta} = \coprod_{k=1}^{\infty} \boldsymbol{\Theta}_k, \text{ where}$$

$$\boldsymbol{\Theta}_k = \Big\{ \theta = \big(\boldsymbol{\mu}_{\cdot,1:k-1}, e_{1:k-1}, \nu_{1:k} \big) \in (\overset{\circ}{\mathbf{S}}_d)^{k-1} \times (0,1)^{k-1} \times (\mathbb{R}^+)^k \; : \; (5) \text{ holds} \Big\}.$$

For $k \geq 1$, we introduce the re-parametrization maps for k-mixtures

$$\Gamma_k : \theta \in \boldsymbol{\Theta}_k \mapsto \big(\boldsymbol{\mu}_{\cdot,\,1:k}, p_{1:k}, \nu_{1:k} \big) \in \Psi_k \, ,$$

which allows to define

$$\begin{aligned} \Gamma : \boldsymbol{\Theta} &\longrightarrow \Psi \\ \theta \in \boldsymbol{\Theta}_k &\longmapsto \Gamma_k(\theta) \in \Psi_k \, . \end{aligned}$$

In the sequel, we denote h_θ a DM density with parameter $\theta \in \boldsymbol{\Theta}$. As opposed to the Ψ-parametrization from Boldi and Davison (2007), we refer to ours as the $\boldsymbol{\Theta}$-parametrization.

3.2. Prior definition

We denote π the prior distribution and also, for the sake of simplicity, the prior density. To prevent numerical issues, *i.e.* to facilitate storage and avoid numerically infinite likelihoods, it appears preferable to restrict the prior's support to a (large) bounded subset

$$\boldsymbol{\Theta}_B = \prod_{k=1}^{k_{\max}} \mathbf{S}_d^{k-1} \times [0, e_{\max}]^{k-1} \times [\nu_{\min}, \nu_{\max}]^k \tag{10}$$

with, typically, $k_{\max} = 15$, $\nu_{\min} = \exp(-2)$, $\nu_{\max} = 5\,10^3$ and $e_{\max} = 1-10^{-6}$. The general definition of $\boldsymbol{\Theta}_B$ allows the case $\boldsymbol{\Theta}_B = \boldsymbol{\Theta}$, by setting the truncation bounds to $e_{\max} = 1, \nu_{\min} = -\infty, \nu_{\max} = +\infty, k_{\max} = +\infty$. Then, the prior can be defined as desired on $\boldsymbol{\Theta}_B$, according to the user's beliefs. Here is described an example of prior specification (the one used in our simulations), allowing the user to control the concentration of the mean vectors $\boldsymbol{\mu}_{\cdot,1:k}$ around the global center of mass $\boldsymbol{\gamma}_0$ (again, *a priori*). Recall that a DM angular density with mean vectors located near the simplex' center, together with high concentration parameters $\nu_{1:k}$, corresponds to high levels of dependence among extreme observations. On the contrary, mean vectors near the vertices or low concentrations are associated with low levels of dependence. As usual, the prior's impact will vanish with large sample sizes, but this kind of control may be useful for small samples, if prior expert knowledge is available.

Conditionally on k, $\boldsymbol{\nu}$ is *a priori* independent from $(\boldsymbol{\mu}, \mathbf{e})$

$$\pi(k,\, \boldsymbol{\mu}, \mathbf{e},\, \boldsymbol{\nu}) = \pi_k(k)\, \pi_{\mu,e}(\boldsymbol{\mu}, \mathbf{e} \mid k)\, \pi_{\boldsymbol{\nu}}(\boldsymbol{\nu} \mid k)\ .$$

The prior π_k is a truncated geometric distribution

$$\pi_k(k) \propto \left(1 - \frac{1}{\lambda}\right)^{k-1} \frac{1}{\lambda} \mathbf{1}_{[1,k_{\max}]}(k)$$

with typical values for λ ranging between 1 and 10. The concentration vector $\boldsymbol{\nu}$ has a truncated multivariate log-normal distribution (denoted logN) with independent components, from which simulation is straightforward. Namely, we set

$$\forall j \in \{1, \ldots, k\},\ \pi_{\nu,j} \propto \mathbf{1}_{[\nu_{\min}, \nu_{\max}]}\ \mathrm{logN}(m_\nu, \sigma_\nu^2). \tag{11}$$

The joint distribution for $\boldsymbol{\nu}$ is the product measure $\pi_{\boldsymbol{\nu}} = \bigotimes_{j=1}^{k} \pi_{\nu,j}$. Finally, the distribution $\pi_{\mu,e}(\cdot \mid k)$ is defined by successive conditioning

$$\pi_{\mu,e}(\boldsymbol{\mu}, \mathbf{e} \mid k) = \prod_{m=1}^{k-1} \pi_{\mu_m}(\boldsymbol{\mu}_{\cdot,m} \mid k,\, \boldsymbol{\mu}_{\cdot,1:m-1}, e_{1:m-1}) \cdots$$
$$\cdots \pi_{e_m}(e_m \mid k,\, \boldsymbol{\mu}_{\cdot,1:m}, e_{1:m-1})$$

where, by convention, $\boldsymbol{\mu}_{\cdot,1:0} = \{\boldsymbol{\gamma}_0\}$ and $e_{1:0} = \emptyset$. In general, one does not want to see the mean vectors rejected on the simplex boundary, where the model is not defined, again to avoid numerical problems such as infinite likelihood values. On

the other hand, it may be of interest to control the dispersion of the k mean vectors $\boldsymbol{\mu}_{\cdot,1:k}$. This is achieved by setting

$$\pi_{\mu_m}(\cdot \mid \boldsymbol{\mu}_{\cdot,1:m-1},\, e_{1:m-1}) = \text{diri}\left(\cdot \mid \boldsymbol{\gamma}_m,\, \frac{\chi_\mu}{\min_{1\le i\le d}\{\gamma_{i,m}\}}\right),$$

where χ_μ is a concentration hyper parameter. Recall that $\boldsymbol{\gamma}_m$ depends on the first $m-1$ components through (4). Thus, for $\chi_\mu \geq 1$, the prior density for $\boldsymbol{\mu}_{\cdot,m}$ is bounded; the larger χ_μ, the more $\boldsymbol{\mu}_{\cdot,m}$ concentrates around the current center of mass $\boldsymbol{\gamma}_{m-1}$. For $0 < \chi_\mu < 1$, the prior is unbounded and the prior mass for $\boldsymbol{\mu}_{\cdot,m}$ is concentrated near the simplex boundaries. In our simulations, χ_μ is set to 1.1. Thus, $\boldsymbol{\mu}_{\cdot,m}$ has relatively flat distribution with bounded density, centered around $\boldsymbol{\gamma}_m$.

Concerning the eccentricity parameters, specifying an identical Beta distribution for each e_m would trigger a bias against the last mixture components: the weights p_m would tend to decrease with m. To avoid this issue, a Beta prior on e_m is defined in such a way, that conditionally to $(\boldsymbol{\mu}_{\cdot,1:m}, e_{1:m-1})$, the expectancy of e_m correspond to a weight ratio p_m/ρ_{m-1} close to $1/(k-m+1)$. Rearranging (9), we have $e_m = \frac{p_m/\rho_{m-1}}{T_m\,(1-p_m/\rho_{m-1})}$. The ideal situation $p_m/\rho_{m-1} = 1/(k-m+1)$ thus corresponds to $e_m = \big(T_m\,(k-m)\big)^{-1}$, which may be greater than one. The distribution's mean is thus set to

$$M_{e,m} = \min\left\{(T_m(k-m))^{-1}\,,\, e_{\text{mean.max}}\right\},$$

where $e_{\text{mean.max}} = 99/100$. Then, another concentration parameter χ_e is introduced and typically set to 1.1. Finally, the Beta parameters $(a_{1,m}, a_{2,m})$ for the mth eccentricity's distribution are set to

$$a_{1,m} = \frac{\chi_e}{\min\{M_{e,m}, 1-M_{e,m}\}}\, M_{e,m}\ ,$$
$$a_{2,m} = \frac{\chi_e}{\min\{M_{e,m}, 1-M_{e,m}\}}(1-M_{e,m}).$$

and $\pi_{e,m}(\cdot \mid k, \boldsymbol{\mu}_{\cdot,1:m}, e_{1:m-1}) \propto \text{beta}\,(\cdot \mid a_{1,m}, a_{2,m})\,\mathbf{1}_{[0,e_{\max})}(\cdot)$ where beta denotes the Beta density.

The Directed acyclic graph in Figure 2 summarizes the model specification. Simulating parameters $(\boldsymbol{\mu}_{\cdot,1:k-1}, e_{1:k-1})$ can be achieved by successively drawing k, then the $\boldsymbol{\mu}_{\cdot,m}$'s and e_m's, in increasing order and finally by using the mapping Γ to obtain $\boldsymbol{\mu}_{\cdot,k}$ and $p_{1:k}$.

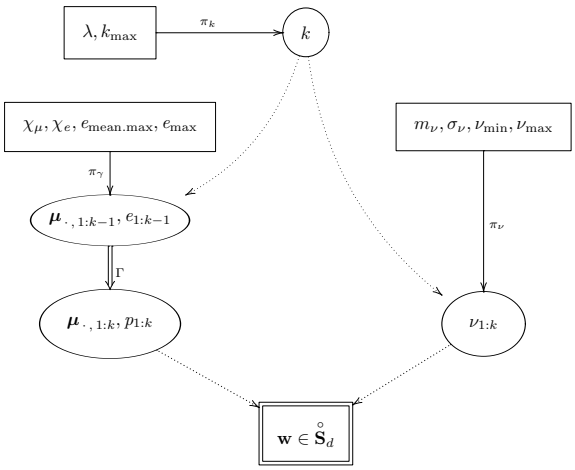

Figure 2: Representation of the conditional dependencies of the DM Bayesian model as a Directed acyclic graph. Hyper-parameters appear in simple square frames, parameters in oval frames and observations in a double square frame. Simple arrows denote probabilistic relations whereas double arrows stand for deterministic ones.

3.3. Model consistency

Boldi and Davison (2007) have shown that the family of finite constrained mixtures of Dirichlet densities is weakly dense in the set of admissible angular measures. Following their steps, we investigate weak consistency properties of the posterior in the re-parametrized model. It is well known (see *e.g.* Freedman, 1963) that weak density does not entail weak consistency, unless some additional regularity assumptions are satisfied, which are detailed in this section. Since the mixture model is not identifiable (several *parameters* θ's correspond to a single density h), we use non-parametric consistency results, which allow one to work with the *densities* themselves. Most of the theoretical background required here may be found in Ghosal et al. (1999) and is derived from Schwartz (1965). For a recent review about available theorems for different types of consistency in the non-parametric case, in particular for the (stronger) Hellinger consistency, the reader may also refer *e.g.* to Walker (2004) and the references therein. Recall that a *weak neighborhood* U of some density h_0 on the sample space $\mathbf{S}_d$ is a family of probability densities containing a finite intersection of subsets of the kind

$$\left\{ h \ : \ \left| \int_{\mathbf{S}_d} \left(h(\mathbf{w}) - h_0(\mathbf{w})\right) g(\mathbf{w}) \ \mathrm{d}\mathbf{w} \right| < \epsilon \right\} \ ,$$

where $\epsilon > 0$ and g is some bounded, continuous function defined on $\mathbf{S}_d$. Similarly, if $(\mathbf{\Theta}, \mathcal{T})$ is a measurable parameter space indexing a family of densities $(h_\theta)_{\theta \in \Theta}$, a weak neighborhood of some $\theta_0 \in \mathbf{\Theta}$ is a weak neighborhood of h_{θ_0} restricted to $\mathbf{\Theta}$ (the weak topology on $\mathbf{\Theta}$ is the trace of the weak topology defined on the densities). Let π be a prior on $\mathcal{T}$ and π_n be the posterior, given independent, identically distributed (*i.i.d.*) observations $\mathbf{W}_1, \ldots, \mathbf{W}_n$, sampled from a probability measure h_0. The posterior is said to be *weakly consistent* at h_0 if, with h_0-probability one, for all weak neighborhood U of h_0, $\pi_n(U^c) \underset{n\to\infty}{\longrightarrow} 0$. It is clear from the definition that two distinct parameters $\theta_1 \neq \theta_2$ defining the same density $h_{\theta_1} = h_{\theta_2}$ will automatically belong to the same weak neighborhoods, so that identifiability is not an issue anymore. Also, weak consistency is usually sufficient for most applications, because the angular density is mainly destined to be integrated against some bounded, continuous function. For example, probabilities of a joint excess of high multivariate thresholds $(u_1, \ldots, u_d)$ are derived by integration of the angular density against $g(\mathbf{w}) = \min(w_1/u_1, \ldots, w_d/u_d)$.

One classical way to prove weak consistency at some density h_0 is to use Schwartz's theorem (Schwartz, 1965, theorem 6.1), which guarantees it under a relatively limited number of assumptions, the most crucial of which being that the prior assign positive mass to any *Kullback-Leibler (KL) neighborhood* of h_0 (see Appendix B for details). We call this property the *KL condition*. Recall that the KL neighborhoods are defined in terms of the KL divergence between two densities, which is the non-negative quantity $KL(h_0, h) = \int_{\mathbf{S}_d} \log(h_0(\mathbf{w})/h(\mathbf{w}))h_0(\mathbf{w})\,\mathrm{d}\mathbf{w}$. A KL neighborhood of some density h_0 is thus a set of densities of the form $K_{h_0,\epsilon} = \{h : KL(h_0, h) < \epsilon\}$, for some $\epsilon > 0$. The *KL support* of the *prior* is the set of all densities for which $\pi(K_{h,\epsilon}) > 0$ for all $\epsilon > 0$. The KL condition is thus that h_0 be in the KL support of the prior. A generally weaker assumption is that h_0 be in the KL *closure* of the model, *i.e.* that any KL neighborhood of h_0, regardless of its prior mass, contain a density h_θ from the model. The KL support is included in the KL closure but the converse may not hold (*e.g.* if the prior does not have full support in the model).

The following proposition (see Appendix B for a proof) establishes the posterior consistency of the re-parametrized DM model on the KL closure of Θ_B for a general class of priors. Here, a 'Euclidean open set' in $\mathbf{\Theta}$ is any union of open sets for the Euclidean topology on the $\mathbf{\Theta}_k$'s. These open sets define the co-product topology induced by the Euclidean topology on the disjoint union $\coprod_k \mathbf{\Theta}_k$.

Proposition 2. *Let π be a prior on the* DM *model assigning positive mass to any non-empty Euclidean open subset of $\mathbf{\Theta}_B$, where $\mathbf{\Theta}_B$ is defined by* (10)*. If h_0 is in the Kullback-Leibler closure of $\mathbf{\Theta}_B$, then the* posterior *is weakly consistent at h_0.*

In particular, for all $\theta_0 \in \mathbf{\Theta}_B$, the posterior is weakly consistent at h_{θ_0}.

In particular, the prior π defined in section 3.2 satisfies the requirement of the statement. This is also the case of the prior defined by Boldi and Davison (2007) on the original model, which can be seen by using the one-to-one mapping Γ. One must note that this result put together with the weak density result from Boldi and Davison is not sufficient to prove weak consistency at *all* angular measure with continuous density on the simplex, even if one takes infinite bounds for $\mathbf{\Theta}_B$, so that $\mathbf{\Theta}_B = \mathbf{\Theta}$. Indeed, the KL topology is thinner than the weak topology, which means that, in general, the KL condition may not be verified even for a density in the weak support of the model. Freedman (1963) provides an example of weakly inconsistent model in a discrete case where the prior still assigns positive mass to all weak neighborhoods of h_0.

For the sake of simplicity we assume in this paper that the true distribution belongs to the model or to its KL closure. However, it would be of interest to investigate the extent of the latter. Also, when the model is 'incorrect' (*i.e.* the KL divergence between the model and the truth is positive), it might be possible to exploit results from Bunke and Milhaud (1998) and show that the posterior concentrates around pseudo-true parameters minimizing the KL divergence between the true h_0 and the model. Bunke and Milhaud (1998)'s results are valid for parametric models containing only bounded densities, so that one should impose a maximum number of mixture components and restrict the model to Dirichlet densities such that $\nu\mu_i \geq 1$ for all $i \in \{1, \ldots, d\}$.

4. Metropolis algorithm

We describe in this section a trans-dimensional Metropolis algorithm to sample the posterior distribution, which we call *Metropolis for Dirichlet mixture*, or, in short, M-DM. It belongs to the class of trans-dimensional (with reversible jumps) *Metropolis within Gibbs* algorithms (MH-Gibbs), as described *e.g.* in Roberts and Rosenthal (2006).

One key principle of the M-DMalgorithm is to use the data to construct the proposal distribution for the mean vectors $\boldsymbol{\mu}_{\cdot,m}$. At each step of the algorithm, three classes of proposal moves are possible: *regular moves*, *trans-dimensional moves* and *shuffle moves*. During a *regular move*, either a mean vector $\boldsymbol{\mu}_{\cdot,m}$, or an eccentricity parameter e_m, or a concentration parameter ν_m is picked out of the current state as a candidate for a move. If a mean vector $\boldsymbol{\mu}_{\cdot,m}$ is chosen, it is thrown back in regions of $\mathbf{S}_d$ where data points concentrate. *Trans-dimensional* moves consist of *split* and *combine* moves. During a *split* (*resp. combine*) *move*, an additional

mixture component is created in the $\boldsymbol{\Theta}$-parametrization. (*resp.* the last component is removed) and the 'last' mean vector $\boldsymbol{\mu}_{\cdot,k} = \boldsymbol{\gamma}_{k-1}$ is adjusted accordingly. Finally, *shuffle move* do not alter the likelihood but are designed to improve the chain's mixing properties. They simply consists in transposing two indices in the Ψ- parametrization and deducing the corresponding $\boldsymbol{\Theta}$-parametrization. They thus correspond to a discrete transition kernel. The probability of choosing a *regular move*, a *trans-dimensional* move or a *shuffle move* have been respectively set to $c_{\text{reg}} = .5$, $c_{\text{trans}} = 1/3$ and $c_{\text{shuf}} = 1/6$.

In the remainder, the proposal variables, the proposal distributions and densities, and the acceptance probability ratios are respectively denoted $(\,\cdot\,)^*$, $Q(\,\cdot\,,\,\cdot^*)$, $q(\,\cdot\,,\,\cdot^*)$, and $r(\,\cdot\,,\,\cdot^*)$; θ_t denotes the chain's state at time (iteration) t. The starting value is generated according to a prior distribution.

4.1. Regular moves

If $\theta_t = \big(\boldsymbol{\mu}_{\cdot,1:k-1}(t), e_{1:k-1}(t), \nu_{1:k}(t)\big) \in \boldsymbol{\Theta}_k$, then $3k-2$ regular moves are possible. Three subclasses are defined: *μ-moves*, *e-moves* or *ν-moves*, depending on the type of component affected. The choice between subclasses is made under equi-probability.

- ν-moves affect one component $\nu_m(t)$ of the concentration vector $\boldsymbol{\nu}$. The proposal density $q_\nu(\nu_m(t), \nu_m^*)$ is log-Normal, with mean parameter $\log(\nu_m(t))$ and standard-deviation parameter typically set to $\log(1+0.5^2)$ (on the log scale).

- Similarly, e-moves affect one eccentricity parameter $e_m(t)$. The proposal density $q_e(e_m(t), e_m^*)$ is a Beta density with mode at $e_m(t)$. The latter is constructed by fixing a *recentring* parameter ϵ_e^* (typically set to 0.2). Then, the Beta parameters are

$$a_1 = \left[\frac{\epsilon_e^*}{2} + (1-\epsilon_e^*).e_m(t)\right]\frac{2}{\epsilon_e^*}\,;\; a_2 = \left[1 - \left(\frac{\epsilon_e^*}{2} + (1-\epsilon_e^*).e_m(t)\right)\right]\frac{2}{\epsilon_e^*}\,.$$

  During an e-move affecting e_m, the weights $p_{m:k}^*$ and the last mean vector $\boldsymbol{\mu}^*_{\cdot,k}$ (in the Ψ-parametrization) are modified according to the mapping $\Gamma : \theta \mapsto \boldsymbol{\psi}$.

- μ-moves affect one of the $k-1$ first mean vectors. Again, the subsequent weights $p_{m:k}^*$ and the last vector $\boldsymbol{\mu}^*_{\cdot,k}$ in $\boldsymbol{\psi}^*$ are modified according to Γ. The proposal $\boldsymbol{\mu}^*_{\cdot,m}$ follows a DM distribution with density $q_\mu(\boldsymbol{\mu}_{\cdot,m}(t),\,\cdot\,)$,

constructed from the angular data $\mathbf{w}_{1:n}$. The mixture is multi-modal, with one mode located at each angular data point, and weights penalizing the distance between the considered data point and the current mean vector $\boldsymbol{\mu}_{\cdot,m}(t)$. The precise construction is a generalization of the e-move distribution. More details are provided in Appendix D.

The acceptance probability for each regular move is classically given by (*e.g.* for e-moves affecting the mth coordinate)

$$r(e_m(t), e_m^*) = \min\left(1, \frac{h_{\theta^*}(\mathbf{w}_{1:n})\pi(\theta^*)}{h_{\theta_t}(\mathbf{w}_{1:n})\pi(\theta_t)} \frac{q_e(e_m^*, e_m(t))}{q_e(e_m(t), e_m^*)}\right).$$

4.2. Trans-dimensional moves

4.2.1. Split moves

This type of move is only proposed when $k < k_{\max}$. A new mean vector $\boldsymbol{\mu}^*_{\cdot,k}$ is generated in a neighborhood of $\boldsymbol{\mu}_{\cdot,k}(t)$, similarly to the proposal rule for the μ-moves, and the last eccentricity parameter e_k^* is proposed according to the prior, see Appendix D.2 for details. Finally, the last mean vector $\boldsymbol{\mu}^*_{\cdot,k+1}$ is deduced from the re-parametrization map Γ.

4.2.2. Combine moves

These deterministic moves are allowed for $k \geq 2$. They simply consist in removing the last component $(\boldsymbol{\mu}_{\cdot,k-1}, e_{k-1}, \nu_k)$ from the Ψ-parametrization. The last mean vector $\mu^*_{\cdot,k}$ in the Ψ- parametrization is thus the center of mass of the two last mean vectors in the current state.

4.2.3. Acceptance ratio for trans-dimensional moves

From Green (1995), the *posterior* distribution is invariant under a trans-dimensional move if we set the acceptance probability, for a split move, to

$$r_{\text{split}} = \min\Big\{1, \frac{h_{\theta^*}(\mathbf{w}_{1:n})\pi(\theta^*)}{h_{\theta_t}(\mathbf{w}_{1:n})\pi(\theta_t)} \frac{p_c(k+1)}{p_s(k)} \times \cdots \\ \left[q_{\mu,\text{split}}(\theta_t, \boldsymbol{\mu}^*_{\cdot,k})\, q_{e,\text{split}}(\theta_t, e_k^* | \boldsymbol{\mu}^*_{\cdot,k})\, q_{\nu,\text{split}}(\theta_t, \nu_{k+1}^*)\right]^{-1}\Big\},$$

and, for a combine move, to

$$r_{\text{combine}} = \min\Big\{1, \frac{h_{\theta^*}(\mathbf{w}_{1:n})\pi(\theta^*)}{h_{\theta_t}(\mathbf{w}_{1:n})\pi(\theta_t)} \frac{p_s(k-1)}{p_c(k)} \times \cdots$$
$$q_{\mu,\text{split}}(\theta^*, \boldsymbol{\mu}_{\cdot,k}(t))\, q_{e,\text{split}}(\theta^*, e_k | \boldsymbol{\mu}_{\cdot,k}(t))\, q_{\nu,\text{split}}(\theta^*, \nu_k(t))\Big\},$$

where $p_c(k)$ and $p_s(k)$ are respectively the probability of choosing a *combine* or a *split* move, when the current state is in $\boldsymbol{\Theta}_k$. Namely, we have set $p_s = \mathbf{1}_{k=1} + \frac{1}{2}\mathbf{1}_{1<k<k_{\max}}$ and $p_c = \mathbf{1}_{k=k_{\max}} + \frac{1}{2}\mathbf{1}_{1<k<k_{\max}}$. Note that the Jacobian appearing in Green's balance condition is, in our case, equal to one. Indeed, the additional component is directly simulated, without further mapping.

4.3. Shuffle moves

These moves do not affect the density h_θ, but improve the convergence of the algorithm. Without shuffling, the weights affected to the last component of the mixture would have a tendency to decrease, as the number of mixture components increases, by a stick breaking effect. Let k be the number of components at step t, $\boldsymbol{\psi}_t = (\boldsymbol{\mu}_{\cdot,1:k}(t), p_{1:k}(t), \nu_{1:k}(t))$. Let $m_1, m_2 \leq k$, and τ_{m_1,m_2} be the transposition between elements indexed by m_1 and m_2 in $\boldsymbol{\psi}_t$. Let $\varphi_{m_1,m_2} = \Gamma^{-1} \circ \tau_{m_1,m_2} \circ \Gamma$. The proposal parameter is then defined by $\theta^* = \varphi_{m_1,m_2}(\theta_t)$. The mapping φ_{m_1,m_2} is differentiable, and we prove in Appendix C that, setting

$$r_{\text{shuffle},m_1,m_2}(\theta_t, \theta^*) = \min\left(1, \frac{h_{\theta^*}(\mathbf{w}_{1:n})\pi(\theta^*)}{h_{\theta_t}(\mathbf{w}_{1:n})\pi(\theta_t)} \left|\text{Jac}(\varphi_{m_1,m_2})_{[\theta_t]}\right|\right)$$

as an acceptance probability for this move, the posterior is invariant under the shuffle kernel. The involved Jacobian is (see Appendix D.3)

$$\left|\text{Jac}(\varphi_{m_1,m_2})_{[\theta_t]}\right| = \prod_{m=1}^{k-1} \frac{\rho_{m-1}\, T_m}{(1+e_m\, T_m)^2} \prod_{m=1}^{k-1} \frac{(1+e_m^*\, T_m^*)^2}{\rho_{m-1}^*\, T_m^*}, \tag{12}$$

where the $e_m^*, \rho_{m-1}^*, T_m^*$'s (*resp.* the e_m, ρ_{m-1}, T_m's) are relative to the proposal parameter $\theta^* = \varphi_{m_1,m_2}(\theta_t)$ (*resp.* θ_t), and the T_m's are defined in Proposition 1.

5. Ergodicity properties of the M-DMalgorithm

There is an abundant literature concerning asymptotic convergence of Markov chains towards their objective distribution, see *e.g.* Meyn et al. (1993) for an extensive exposition. In short, let $\tilde{\pi}$ is an objective probability on $(\Theta, \mathcal{T})$, *i.e.* a

distribution from which one wishes to generate a sample (here, $\tilde{\pi}$ is the *posterior* π_n and $\Theta = \mathbf{\Theta}_B$). Let $\tilde{\pi}$'s density with respect to some reference measure $\mathrm{d}\eta$ be known up to a normalizing constant. We also denote $\tilde{\pi}$ this un-normalized density.We shall use a classical result (see *e.g.* Rosenthal, 2001; Roberts and Rosenthal, 2006; Tierney, 1994): under regularity assumptions, if an aperiodic Markov chain is generated by a transition kernel $K(\theta, \cdot)$ admitting $\tilde{\pi}$ as an invariant probability measure, and if $K(\theta, \cdot)$ is η-irreducible, then for $\tilde{\pi}$-almost all starting value, the law $K^n(\theta_{\text{start}}, \cdot)$ defined by the n-step transition kernel converges in total variation distance towards $\tilde{\pi}$.

The regularity assumption is that $\mathcal{T}$ be countably generated. This is not too restrictive, since it is true in any case where Θ is some Borel space and $\mathcal{T}$ is its Borel σ-field. In particular, this is true in our context, since $\mathbf{\Theta}$ can be identified with a finite union of open subsets in finite dimensional euclidean spaces. *Aperiodicity* means the state space cannot be finitely partitioned into subsets $\Theta_1, \ldots, \Theta_d$ $(d > 1)$ such that for $1 \leq i < d$ and $\theta_i \in \Theta_i$, $K(\theta_i, \Theta_{i+1}) = 1$, and for $\theta_d \in \Theta_d$, $K(\theta_d, \Theta_1) = 1$. Also, $\tilde{\pi}$ is *invariant* by K if $\forall \theta \in \Theta, \forall A \in \mathcal{T}$, $\int_\Theta K(\theta, A)\ \mathrm{d}\tilde{\pi}(x) = \tilde{\pi}(A)$. Such a $\tilde{\pi}$ is also called *stationary*. Finally, *η-irreducibility* stipulates that for all set $A \subset \Theta$ such that $\eta(A) > 0$, for all $\theta \in \Theta$, for some $t \in \mathbb{N}$, $K^t(\theta_{\text{start}}, A) > 0$.

Convergence in total variation distance entails a mean ergodicity property that can be used in conjunction with weak consistency. Namely, for all $\tilde{\pi}$ integrable function g, and for $\tilde{\pi}$-almost all starting value, it implies

$$\frac{1}{T}\sum_{t=1}^{T} g(\theta_t) \underset{T\to\infty}{\longrightarrow} \mathbb{E}_{\tilde{\pi}}(g), \quad P_{\theta_{\text{start}}} \text{ almost surely,} \tag{13}$$

where $P_{\theta_{\text{start}}}$ represents the probability measure on $\left(\mathbf{\Theta}^{\mathbb{N}}, \mathcal{T}^{\otimes\mathbb{N}}\right)$ induced by the Markov kernel and the initial state θ_{start}, and θ_t is the random state at time t. Note that, from Roberts and Rosenthal (2004) (*cf* their remark following Corollary 6), aperiodicity is not required for (13). In our case, a natural choice for η is the Lebesgue measure on the Euclidean co-product space $\mathbf{\Theta}_B$, defined by (10). In order to verify that (13) holds for the M-DMalgorithm, we show in Appendix C the following

Proposition 3. *The* M-DM*algorithm generates a η-irreducible, aperiodic Markov chain admitting the posterior π_n as an invariant probability measure.*

The original part of the proof of Proposition 3 concerns the invariance of the discrete shuffling kernel. Indeed, standard reversibility arguments are only valid for continuous proposal kernels. In contrast, irreducibility and aperiodicity are verified in a classical way and some ideas are in common *e.g.* with Roberts and

Smith (1994) (in the context of the standard Gibbs sampler) and Guillotte et al. (2011) (pp. 392-393, proofs 6.3.2 and 6.3.3, together with their Appendix A.5, for a particular trans-dimensional Gibbs sampler). As noted by the latter authors, the literature is scarce concerning general conditions for irreducibility and aperiodicity in a trans-dimensional context. We thus provide a proof that suits our purposes.

The $\tilde{\pi}$-null set on which (13) is not guaranteed may be problematic because its extent is unknown. If, in addition to the properties listed in Proposition 3, a Markov chain is Harris recurrent, then the result holds for *all* starting value. A η-irreducible Markov chain with stationary distribution $\tilde{\pi}$ is said *Harris-recurrent* if for all $A \subset \Theta$, such that $\eta(A) > 0$, the stopping time $\tau_A = \inf\{N \geq 1 : \theta_N \in A\}$ is almost surely finite for all starting value: $P_{\theta_{\text{start}}}(\tau_A < \infty) = 1$ for all θ_{start}. Full-dimensional MH algorithms are Harris-recurrent under weak assumptions regarding the support of the proposal distributions. A short and self contained proof was recently proposed by Asmussen and Glynn (2010), see also *e.g.* Rosenthal (2001); Roberts and Rosenthal (2004) or Roberts and Rosenthal (2006) for a review of the properties of the class of MH-Gibbs and trans-dimensional MH algorithms. Harris-recurrence is less easily achieved for the two latter classes than for the full-dimensional MH algorithm, and the question is even stated as an open problem in the case of coordinate mixing, trans-dimensional Markov chains (which is precisely our framework, see paragraph 'shuffle moves' in the preceding section). Similarly to Guillotte et al. (2011), we do not prove Harris-recurrence for the M-DMalgorithm. In our case, the difficulty comes from discontinuities of the proposal density around singular points where (5) does not hold. However, generating the starting value according to the prior and noticing that $\pi \ll \tilde{\pi}$, the starting value will almost-surely not belong to the problematic set.

We now turn to practical implications of (13) (which itself derives from Proposition 3). As discussed in section 3.3, for applied purpose, the quantity of interest is often obtained as an integral of some bounded, continuous function g defined on the simplex, with respect to the angular measure H. We thus define, for such a g,

$$\begin{aligned} \tilde{g}(\theta) &= \int_{\mathbf{S}_d} g(\mathbf{w}) h_\theta(\mathbf{w}) \, \mathrm{d}\mathbf{w} \\ &:= \langle g, h_\theta \rangle \, . \end{aligned} \tag{14}$$

The function $\tilde{g}$ is bounded by $\|g\|_\infty$, and its continuity (for the weak topology) may be verified: The arguments are the same as those leading to the continuity of κ, in the proof of Proposition 2. Note that standard arguments involving the continuity of the inner product cannot be used instead, because h_θ does not belong to the L^2 space corresponding to the inner product (*i.e.* h_θ^2 is not integrable) if

one Dirichlet exponent $\nu_m \mu_{j,m}$ is less than 1/2.

As a consequence of the weak continuity, provided that the true measure h_0 satisfies the assumptions of Proposition 2 (so that the posterior is weakly consistent at h_0), we have

$$\mathbb{E}_{\pi_n}(\tilde{g}) \underset{n\to\infty}{\longrightarrow} \tilde{g}(h_0) = \langle g, h_0 \rangle \qquad (h_0\text{-a.s})\,.$$

Combining this with (13) shows that

$$\lim_{n\to\infty} \left(\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \langle g, h_{\theta_t^n} \rangle \right) = \langle g, h_0 \rangle \qquad (h_0 \times P_{\theta_{\text{start}}}\text{-a.s.})\,, \tag{15}$$

where n is the data sample size and θ_t^n is the current state at time t of the algorithm.

6. Convergence assessment

6.1. Choice of the monitored quantity

In this section, we propose a method to assess goodness-of-fit and monitor MCMC convergence, *i.e.* to verify in practice that the double limit in (15) has approximately been reached. The method also allows to check that the mixing properties of the generated chains are good enough to provide a representative sample from the posterior. Non-identifiability and shuffling prevent from monitoring the parameter components generated by the algorithm. On the other hand, there is no obvious way to visualize the evolution of the generated densities $(h_{\theta_t})_t$ themselves. One solution is to extract suitable numerical quantities that represent the generated densities, in relation to (15), and then to apply standard convergence tests to the numerical representations. For example, in the bi-variate case, Boldi and Davison (2007) monitor the evolution of the dependence measure corresponding to the density h_{θ_t}: $\tilde{g}(\theta_t) = \int_0^1 \min(w, 1-w) h_{\theta_t}(w)\,\mathrm{d}w$. This quantity has an analytical expression (using incomplete Beta functions) in the case $d = 2$ only.

The ideas developed here aim at proposing suitable g's, for which $\tilde{g}(\theta) = \langle g, h_\theta \rangle$ in (15) can easily be derived in arbitrary dimension, and such that the M-DMestimates $\frac{1}{T}\sum_{t=1}^{T} \langle g, h_{\theta_t} \rangle$ can be compared to a reference value (the true value $\langle g, h_0 \rangle$ for simulations or an empirical estimate in realistic cases). For this purpose, it is very convenient to choose g in the set of bounded Dirichlet distributions, which are those with parameters $(\boldsymbol{\mu}, \nu)$ verifying $\nu\,\mu_i \geq 1$, for all $i = 1, \ldots, d$. To see this, suppose that h and g are two Dirichlet densities with respective parameters (μ, ν) and $(\tilde{\mu}, \tilde{\nu})$, and suppose that g is bounded, so that $\tilde{\nu}\,\tilde{\mu}_i \geq 1$ for all $i \leq d$. Then, direct calculations yield the (rather complicated, but tractable) expression

$$
\begin{aligned}
\langle g, h\rangle &= \int_{\mathbf{S}_d} g(\mathbf{w})h(\mathbf{w})\ \mathrm{d}\mathbf{w} \\
&= \frac{\Gamma(\nu)\Gamma(\tilde{\nu})}{\prod_{i=1}^d \Gamma(\mu_i\nu)\Gamma(\tilde{\mu}_i\tilde{\nu})} \int_{\mathbf{S}_d} \prod_{i=1}^d w_i^{(\mu_i\nu+\tilde{\mu}_i\tilde{\nu}-1)-1}\ \mathrm{d}\mathbf{w} \\
&= \frac{\Gamma(\nu)\Gamma(\tilde{\nu})}{\prod_{i=1}^d \Gamma(\mu_i\nu)\Gamma(\tilde{\mu}_i\tilde{\nu})} \frac{\prod_{i=1}^d \Gamma(\mu_i'\nu')}{\Gamma(\nu')} \\
&:= \mathcal{I}_{\boldsymbol{\mu},\nu}(\tilde{\boldsymbol{\mu}}, \tilde{\nu})
\end{aligned} \tag{16}
$$

where $\nu' = \nu + \tilde{\nu} - d > 0$ and $\mu_i' = (\mu_i\nu + \tilde{\mu}_i\tilde{\nu} - 1)/\nu'$.

In experiments with simulated data, the true h_0 may be a Dirichlet mixture, in which case the reference $\langle g, h_0\rangle$ has a similar expression. Indeed, there is no further difficulty if the simple Dirichlet h in (16) is replaced with any DM density $h_0 = h_\theta$ with $\theta = (\mathbf{p}, \boldsymbol{\mu}, \nu)$. The quantity $\langle g, \theta\rangle := \langle g, h_\theta\rangle$ is then obtained as a convex combination of $\mathcal{I}_{\boldsymbol{\mu}_{\cdot,m},\nu_m}(\tilde{\boldsymbol{\mu}}, \tilde{\nu})$ with weight vector $\mathbf{p}$ (see (E.1) in Appendix).

When h_0 is unknown, an empirical mean estimator may be used instead: Consider a function g and a data set $\mathbf{W}_{1:n}$ as above. Then, note that $\langle g, h_0\rangle = \int_{\mathbf{S}_d} g\, h = \mathbb{E}_{h_0}(g)$, so that a classical non-parametric estimate of $\langle g, h_0\rangle$ is

$$
\hat{g}^{\mathrm{nonP}} \ = \ \frac{1}{n}\sum_{j=1}^n g(\mathbf{W}_j). \tag{17}
$$

In addition to a reference value, a reference error is needed. It is obtained as the standard deviation $\delta^{\mathrm{nonP}}(g)$ (under h_0) of the estimator $\hat{g}_n^{\mathrm{nonP}}$

$$
\delta^{\mathrm{nonP}}(g) = \frac{1}{\sqrt{n}}\left[\mathrm{Var}_{h_0}(g)\right]^{1/2} = \frac{1}{\sqrt{n}}\left[\mathbb{E}_{h_0}(g^2) - (\mathbb{E}_{h_0}(g))^2\right]^{1/2}, \tag{18}
$$

A closed form when h_0 is a Dirichlet mixture is derived in Appendix E.2. Again, a non parametric estimate is readily available: $\hat{\delta}^{\mathrm{nonP}}(g) = \frac{1}{\sqrt{n}}\left[(\hat{g^2})^{\mathrm{nonP}} - \left(\hat{g}^{\mathrm{nonP}}\right)^2\right]^{1/2}$.

The Dirichlet test functions g's can be interpreted from a statistical point of view, other than being a convenient computational tool. Take g as a highly peaked Dirichlet (*i.e.* with large concentration ν), with mean vector $\boldsymbol{\mu} \in \mathbf{S}_d$. Then $\langle g, h_0\rangle$ is close to $h_0(\boldsymbol{\mu})$ and the $\langle g, h_{\theta_t}\rangle$'s are close to $h_{\theta_t}(\boldsymbol{\mu})$. Thus, closeness of the estimate $\frac{1}{T}\sum\langle g, h_{\theta_t}\rangle$ to the true value may be reformulated in terms of goodness-of-fit of the posterior predictive density in a neighborhood of the simplex point

$\boldsymbol{\mu}$. In practice, this allows to check that the posterior predictive behaves well in regions of interest (for example, in the regions where the observed angular data concentrate). Thus, in this paper, the mean vectors for the Dirichlet test functions g's are drawn in the neighborhoods of the angular data points. More details are gathered in Appendix E.1.

6.2. Assessing convergence in practice

For each case study, the M-DMalgorithm is ran J times (typically, $J = 4$ or $J = 8$) with starting values generated from the prior. A set of Dirichlet test functions $\{g_\ell, 1 \leq \ell \leq L\}$ is randomly chosen from the data and convergence of the jth chain $(\theta_t(j))_{t\geq 0}$ is monitored *via* the mapped chain $\left(\langle h_{\theta_t(j)}, g_\ell\rangle\right)_t$. For the sake of simplicity, we use the convergence assessment tools available in R package coda. First, the stationarity of single mapped chains is investigated using the the Heidelberger and Welch criterion, (Heidelberger and Welch, 1983). The latter is based on a Cramer-von-Mises statistic and is implemented in R function heidel.diag. Under the null hypothesis that the chain has reached its stationary domain, the statistic has standard normal distribution. In a second step, only the stationary chains are retained, and it must be checked that starting values have lost their influence. For such a purpose, we use the diagnostic proposed by Gelman and Rubin (1992) and implemented in R functions gelman.diag and gelman.plot. The principle is to compare within-chain and inter-chain variances. The multivariate Gelman ratio statistic R_G (shrink factor for the L-variate chains) converges to 1 under the null-hypothesis and a typical requirement is that $R_G < 1.1$.

Beside stationarity and mixing properties, goodness-of-fit (*i.e.* accuracy of the density estimates) is of primarily interest. Let h_0 be the 'true' density and consider a test function g. Discarding the first T_1 iterations of each run and considering the sub-samples obtained between iterations $T_1 + 1$ and T_2 ($T_2 > T_1$), the estimate of $\langle g, h_0\rangle$ produced by the M-DMalgorithm, using the $J' \leq J$ stationary chains,

$$\hat{g} = \frac{1}{J'(T_2 - T_1)} \sum_{j=1}^{J'} \sum_{t=T_1+1}^{T_2} \langle g, h_{\theta_t(j)}\rangle \ .$$

Each term of the summation has analytical expression derived from (16). If h_0 is belongs to the model (*e.g.* for a simulation experiment), the true value $\langle g, h_0\rangle$ is known and the *exact DM error* is then

$$\Delta(g) = |\hat{g} - \langle g, h_0\rangle| \ .$$

As a summary, the error ratio

$$r(g) = \frac{\Delta(g)}{\delta^{\text{nonP}}} \tag{19}$$

may be used as a goodness-of fit indicator for the posterior mean estimates. Values lower than one indicate that the DM estimate (for a given test function) is in the expected range of the empirical estimator. If h_0 is unknown, goodness-of-fit may still be assessed by comparing the model estimate with its empirical counterpart, *i.e.* by replacing $\langle g, h_0 \rangle$ with $\hat{g}^{\text{nonP}}$ in (19) and δ^{nonP} with its estimate $\hat{\delta}^{\text{nonP}}$. This defines the *empirical DM error* $\hat{\Delta}(g)$ and the empirical error ratio $\hat{r}(g) = \frac{\hat{\Delta}(g)}{\hat{\delta}^{\text{nonP}}(g)}$.

7. Results

In this section, the re-parametrized algorithm is tested on a variety of simulated data sets, and on the air quality data set recorded in Leeds, presented in the introduction. Comparison is made with the original version of the algorithm. For each data set, five Dirichlet test functions are randomly chosen. Only the chains for which the minimum Heidelberger p-value (over the five test functions) is greater than 0.01 are kept for further analysis. Then, the quality of convergence is measured in terms of number J' of stationary chains, and of mean Heidelberger and Welches p-value $\overline{hw}$ (over the stationary chains and the five test functions). Too low values indicate a lack of stationarity. The multivariate Gelman ratio R_G summarizes the mixing properties, and goodness-of-fit is assessed using the mean error ratio over the five test functions, computed over the stationary chains, $\bar{r} = \frac{1}{5}\sum_{\ell=1}^{5} r(g_\ell)$, as well as the minimum and maximum ratios $r_{\min} = \min_\ell r(g_\ell)$, $r_{\max} = \max_\ell r(g_\ell)$. For these error ratios, lower values indicate better fit.

7.1. Example: tri-variate simulated data

In this example, a sample of one hundred tri-variate points is simulated from a three component DM distribution with parameter $\theta_0 = (\boldsymbol{\mu}_0, p_0, \nu_0)$, with

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 0.3 & 0.2 & 0.475 \\ 0.6 & 0.1 & 0.175 \\ 0.1 & 0.7 & 0.35 \end{pmatrix}, \tag{20}$$
$$p_0 = (5/12, 1/4, 0.5, 1/3), \text{ and } \nu_0 = (15, 11, 20).$$

Figure 3 compares the true density with the posterior predictive resulting from one chain produced by the re-parametrized M-DMalgorithm. To save computational time, only one out of 100 iterations were kept to compute the predictive density.

For the other tests based on integration against Dirichlet densities, the thinning interval was set to 10. The predictive angular density appears to reproduce well

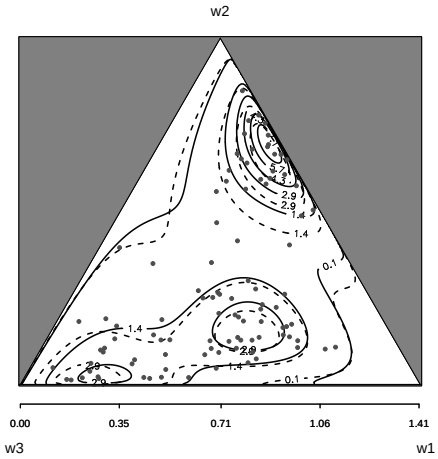

Figure 3: Predictive angular density contours (solid lines) obtained *via* the M-DMalgorithm, on the two-dimensional simplex $\mathbf{S}_3$, inferred with 100 simulated points (Gray points) simulated from the true density (dotted lines) defined by (20).

the characteristics of the mixture. The contour lines of the predictive density obtained with the original version of the DM model are very similar (not shown). However, this visual check is not sufficient to assess the convergence of the chains. For this purpose, we follow the procedure described in section 6.2. Four parallel chains of 50 000 iterations are run, using the re-parametrized algorithm and the original one. The first 10 000 iterations of each chain are discarded. Table 1 summarizes the convergence statistics introduced at the beginning of section 7. Both algorithms perform well in terms of goodness-of-fit, as indicated by the error ratios $\bar{r}, r_{\min}, r_{\max}$. For this data set, the original algorithm even yields better estimates (after averaging over the different parallel chains). Also, in both cases, all the chains are deemed stationary in terms of Heidelberger statistic. However, in terms of mixing properties, summarized by the Gelman shrink factor R_G, the original algorithm is outperformed by the re-parametrized version.

For a more immediate convergence diagnostic, Figure 4 shows the evolution of the quantities $\langle g, h_{\theta_t(j)} \rangle$ (as defined in (14)), where $j \in \{1, \ldots, 4\}$ is the chain index, and of the mean estimates $\hat{g}_{T,j} = \frac{1}{T} \sum_{t \leq T} \langle g, h_{\theta_t(j)} \rangle$, for one given test

	J'	$\overline{hw}$	R_G	$\bar{r}$	$(r_{\min}, r_{\max})$
Re-parametrized	4	0.40	1.01	0.52	$(0.02, 1.05)$
Original	4	0.64	1.37	0.36	$(0.02, 0.72)$

Table 1: Convergence assessment on tri-variate simulated data: comparison between the re-parametrized algorithm (first line) and the original version (second line). From left to right: number of stationary chains, mean Heidelberger p-values, multivariate Gelman shrink factor, mean error ratio (minimum and maximum values), see section 6.2 for details.

function g. Clearly, the mixing properties differ between the two algorithms, so that the original one should be ran with a larger number of iterations to span properly the support of the posterior.

7.2. Example: simulated five-dimensional data

We now turn to five dimensional problems. A 100-points data set is simulated from a four-components DM distribution with parameters

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 0.1 & 0.5 & 0.2 & 0.18 \\ 0.1 & 0.2 & 0.2 & 0.24 \\ 0.1 & 0.1 & 01 & 0.3 \\ 0.2 & 0.1 & 0.3 & 0.18 \\ 0.5 & 0.1 & 0.2 & 0.1 \end{pmatrix},$$

$$p_0 = (0.2, 0.1, 0.2, 0.5), \quad \nu_0 = (30, 40, 20, 25).$$

Four parallel chains of length 200×10^3 are run in each model, from which the first 80×10^3 are discarded. The same convergence diagnostic is performed as for the three dimensional case, results are gathered in Table 2. Visually, the chains in both versions of the algorithm evolve in a very similar way as in Figure 4. The same conclusion can be drawn as in the tri-variate case. The only difference is the number of simulations required to obtain good convergence statistics with the M-DMalgorithm. The computational burden remains reasonable: the typical run-time is of five minutes for one chain.

	J'	$\overline{hw}$	R_G	$\bar{r}$ $(r_{\min}, r_{\max})$
Re-parametrized	2	0.25	1.02	0.59 $(0.06, 1.41)$
Original	3	0.27	2.06	0.87 $(0.26, 1.74)$

Table 2: Convergence assessment on five-variate simulated data, with the same statistics as in Table 1

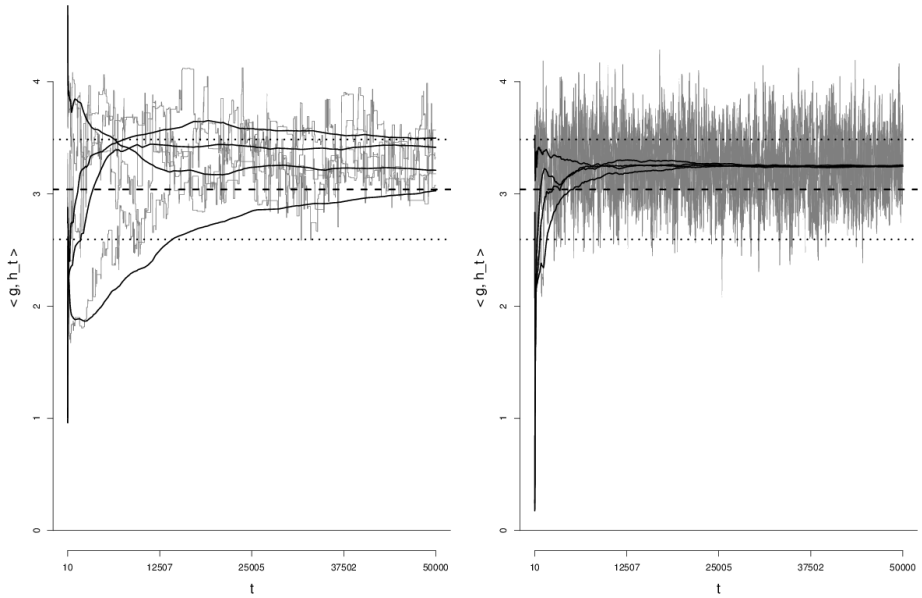

Figure 4: Convergence monitoring with three-dimensional data, with four parallel chains in each model. Integration of the densities generated by the original DM model (left panel) and by the re-parametrized version (right panel) against a Dirichlet density with parameter $\nu\ \boldsymbol{\mu} \simeq (7.67, 2.72, 4.60)$.
Gray lines: Evolution of $\langle g, h_{\theta_t(j)} \rangle$. Black, solid lines: cumulative mean. Dashed line: true value $\langle g, h_0 \rangle$. Dotted lines: true value +/- 1 theoretical standard deviation δ_n^{nonP} of the empirical mean estimate with $n = 100$ points.

One practical implication of the slow mixing on the original parametrization is that posterior credible intervals are difficult to estimate. As an example, Figure 5 displays, for the two parametrizations, the estimated posterior mean of the bivariate angular density for the coordinates pair $(3, 5)$, obtained by marginalization of the five-variate estimated density. The posterior credible band corresponds to the point-wise $0.05 - 0.95$ quantiles of the density. In both cases, the estimates are obtained from the last 120.10^3 iterations of one chain. The estimated credible band with the original algorithm is much thinner than it is with the re-parametrized one. As a consequence, the true density is out of the interval for a large proportion of angular points in $(0, 1)$.

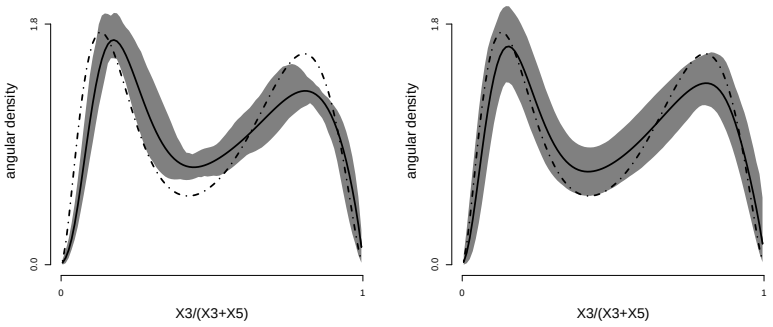

Figure 5: Simulated five-dimensional data (100 points): Bi-variate angular posterior predictive densities for the pair $(3, 5)$. Left panel: Original algorithm; Right panel: re-parametrized version. Dash-dotted line: true density; solid line: posterior predictive; Gray area: posterior credible set at level 0.9.

7.3. Case study: Leeds data set

This data set gathers daily maximum concentrations of five air pollutants: particulate matter (PM10), nitrogen oxide (NO), nitrogen dioxide (NO2), ozone (O3), and sulfur dioxide (SO2). As noted *e.g.* by Heffernan and Tawn (2004), the time series exhibits a daily cycle and short term temporal dependence, so that daily maxima may be considered as independent in time. Following Cooley et al. (2010), marginal distributions are estimated by fitting a generalized Pareto distribution to the upper 0.7 quantile and using the empirical distribution for the remaining observations. Marginal transformation into unit Fréchet is then performed by probability integral mapping. The 100 largest observations (for the L^1 norm) over the 498 non missing five-variate observations are retained for model inference.

For those extremes, the convergence is slow. This may be due to the weak dependence at asymptotic levels found by Heffernan and Tawn (2004), which entails a concentration of the angular points near the boundaries of the simplex, so that the estimated densities are often unbounded. Indeed, when $\nu_m \, \mu_{i,m} < 1$ for some (i, m), the Dirichlet mixture density grows to infinity in the ith vertice and the likelihood is very sensitive to small perturbations of $\mu_{i,m}$. Eight chains of 10^6 iterations each were generated, the first $4\,10^5$ iterations being discarded as a burn-in period. For this data, convergence was slightly enhanced by modifying some of the hyper-parameters for the prior and of the MCMC tuning parameters: the maximum eccentricity $e_{\max}$ was set to $1 - 10^{-3}$, while the maximum expectancy

$e_{\text{mean.max}}$ for the corresponding Beta distribution for the e_m's was set to 0.9. (instead of, respectively, $1 - 10^{-6}$ and 0.99). As for the MCMC tuning parameters, the recentring parameters $\epsilon_\mu^{\text{split}}$ for split-moves and ϵ_e for e-moves are respectively set to 0.3 and 0.4 (instead of 0.5 and 0.2). Results are gathered in Table 3. Here, the error ratio are computed using the empirical estimates $\hat{g}^{\text{nonP}}$ as a reference. Again, mixing remains acceptable in the re-parametrized DM model, provided the run length is long enough, contrary to the original version. Figure 6 shows the projection of the predictive density on three out of the ten two-dimensional simplex faces. This example allows to verify that our estimates are close to those found by Boldi and Davison (2007) using a non-Bayesian EM algorithm. Again, the mean estimates obtained with the original MCMC algorithm are very similar but the posterior $0.05 - 0.95$ quantiles are thinner (not shown).

	J'	$\overline{hw}$	R_G	$\bar{r}$	$(r_{\min}, r_{\max})$
Re-parametrized	2	0.19	1.11	0.64	(0.05, 1.09)
Original	4	0.19	1.66	0.77	(0.12, 1.39)

Table 3: Convergence assessment on Leeds air quality data set, with the same statistics as in Table 1.

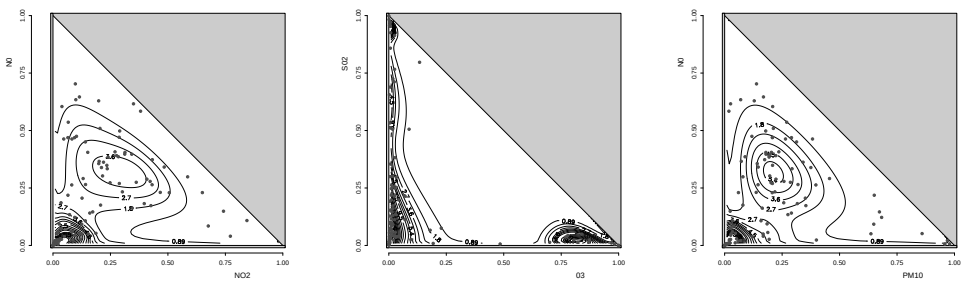

Figure 6: Five dimensional Leeds data set: posterior predictive density. Black lines: projections of the predictive angular density defined on the four-dimensional simplex $\mathbf{S}_5$ onto the two-dimensional faces. Gray dots: projections of the 100 points with greatest L^1 norm.

7.4. Prior influence

In this section, the influence of the prior specification is investigated. The re-parametrized model is fitted on the same simulated five-dimensional data set as in section 7.2, with different values for the hyper-parameters $\lambda, \sigma_\nu, \chi_\mu, \chi_e$ defined in section 3.2. Also, we verify that defining the prior distribution of $(\boldsymbol{\mu}, \mathbf{e})$ jointly, as

in section 3.2, leads to a substantially more reliable inference than when the $\boldsymbol{\mu}_{\cdot,j}$'s and the e_j's are a priori mutually independent. An alternative prior for $(\boldsymbol{\mu}, \mathbf{e})$ is thus defined so that all the mean vectors (*resp.* eccentricities) are independent and uniformly distributed on the simplex (*resp.* the segment $[0, e_{\max}]$). For this simplified prior, the shape hyper-parameter σ_ν is varied in the same way as in the preceding setting.

The default hyper-parameter values are set to

$$\lambda = 5\,,\ k_{\max} = 15\,,$$
$$m_\nu = \log(60)\,, \quad \sigma_\nu^2 = \log(1+5^2)\,, \quad \log(\nu_{\min}) = -2\,, \quad \log(\nu_{\max}) = 5000\,,$$
$$\chi_e = 1.1\,, \quad e_{\text{mean.max}} = 0.99 \quad e_{\max} = 1 - 10^{-6} \quad \chi_\mu = 1.1\,.$$

Starting from this, the hyper-parameters $\lambda, \sigma_\nu, \chi_\mu, \chi_e$ are perturbed, one at a time, see Figure 7 for details. For each hyper-parameters value, four chains are run in parallel, with a burn-in period of 80×10^3 followed by another period of 80×10^3 iterations. The same Dirichlet test functions as in section 7.2 are chosen. Goodness-of-fit is assessed in terms of the average error ratio $\bar{r}^{\mathrm{DM}} = \frac{1}{5}\sum_{\ell=1}^{5} r^{\mathrm{DM}}(g_\ell)$ (left panel of Figure 7) and mixing is checked *via* the multivariate Gelman ratio (right panel) computed on the stationary chains only. On both panels, lower values indicate better properties.

When $\boldsymbol{\mu}$ and e are *a priori* dependent, as in section 3.2, convergence and goodness-of-fit are rather robust to hyper-parameters specification: First, the hyper-parameter λ ruling the number of components has a limited impact, only the value $\lambda = 1$ (which penalizes sharply the number of mixture components) damages the goodness-of-fit. The number of mixture components does not explode for large values of λ (Figure 8), which matches the findings of Boldi and Davison (2007) with the original algorithm. The scores are also approximately constant over the studied range of the other hyper-parameters (Figure 7). Only the large value $\chi_\mu = 8$ damages the mixing properties of the algorithm. The only case of instability is observed with the simplified version of the prior on (μ, e), for which the mixing properties are generally poor. Note that the third Gelman ratio (corresponding to $\sigma_\nu^2 = \log(1+2^2)$) is missing, because less than two chains passed the Heidelberger test for this particular experiment. As a conclusion, the structure of the prior defined in section 3.2 appears to result in mean estimates that are relatively robust to the hyper-parameters specification, and to ensure better mixing properties than the simplified version where $\boldsymbol{\mu}$ and $\mathbf{e}$ are *a priori* independent.

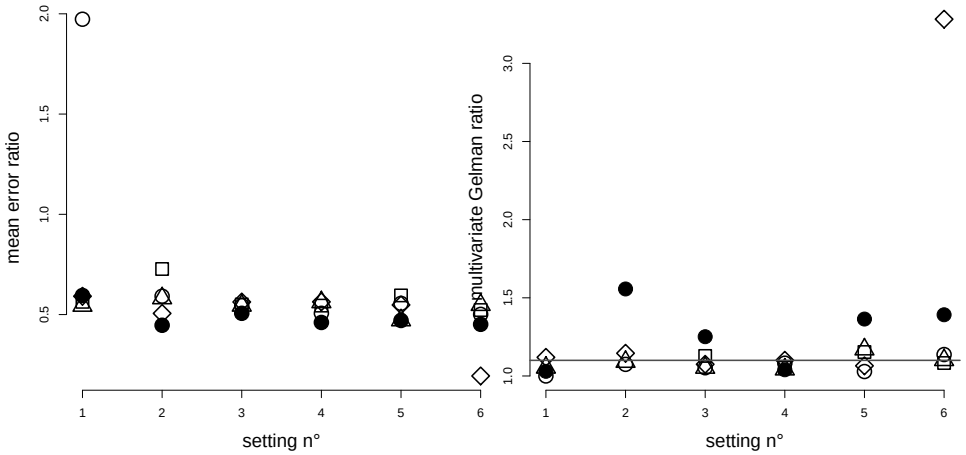

Figure 7: Influence of the prior on the quality of the fit (left panel, mean error ratio $\bar{r}^{\mathrm{DM}}$), and on the chains' mixing properties (right panel, multivariate potential scale reduction factor R_G). •, simplified prior on $(\boldsymbol{\mu}, e)$, influence of σ_ν (variance of the shapes), $\sigma_\nu^2 \in \{\log(1+(0.5)^2), \log(1+1^2), \log(1+2^2), \log(1+5^2), \log(1+10^2), \log(1+20^2)\}$ (from left to right) ; □, dependent prior on $(\boldsymbol{\mu}, e)$, influence of σ_ν, same values for σ_ν; ∘, influence of λ (mean number of mixture components), $\lambda \in \{1, 3, 5, 7, 10, 12\}$; ⋄, influence of χ_μ (concentration of mean vectors), $\chi_\mu \in \{0.5, 1, 1.5, 2, 4, 8\}$; △, influence of χ_e (concentration of eccentricities), $\chi_e \in \{0.5, 1, 1.1, 1.5, 3, 6\}$; Horizontal gray line (right panel), level 1.1 for the Gelman ratio.

7.5. Comparison with other methods for bi-variate data

Here, the M-DMalgorithm is compared with other Bayesian models that have already been proposed for the bi-variate case. Namely, comparison is made with the original DM model and with the non-parametric Bayesian model for bi-variate spectral measure from Guillotte et al. (2011). In the latter, the angular measure is obtained as a smoothed version of a discrete distribution on $(0, 1)$, allowing for atomic masses on $\{0\}$ and $\{1\}$ and satisfying the moments constraint. The parameters' randomness concerns the number and positions of the atoms on $(0, 1)$ defining the underlying discrete distribution (to be smoothed), as well as the amount of mass to be attributed to the boundary.

A simulation study is performed following the same pattern as in Guillotte et al. (2011) and Einmahl and Segers (2009). Bi-variate data sets are simulated from three multivariate extreme value distributions belonging respectively to the Logistic model, to the Asymmetric Logistic model and to the DM model itself (see Appendix F for details). These 'true' distributions are respectively denoted $H_{\mathrm{L}}, H_{\mathrm{AL}}, H_{\mathrm{DM}}$. Contrary to the two other ones, the Asymmetric logistic distri-

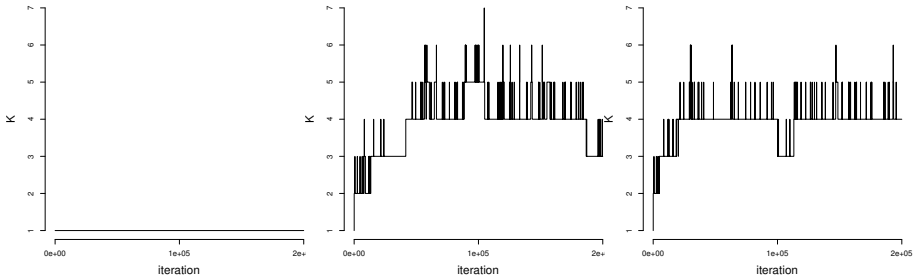

Figure 8: Influence of hyper-parameter λ on the number k of mixture components generated by the reversible-jump sampler. From left to right, evolution of k in the MCMC run, for $\lambda \in \{1, 3, 12\}$, and a 100-points data set issues from four-components Dirichlet mixture.

bution has point masses at 0 and 1. For each H_m ($m \in \{\mathrm{L}, \mathrm{AL}, \mathrm{DM}\}$), 100 data sets of size 1000 are simulated and the three Bayesian models are fitted. Following Guillotte et al. (2011), for the non-parametric Bayesian model, a bi-variate threshold $\mathbf{u} = (u, u)$ is chosen, with u equal to the theoretical 0.9 marginal quantile, and the original algorithm is modified so that the marginal parameters are set to their true values. The number of angular observations retained for fitting both versions of the DM model is the same as the number of points in the upper square region $[u, \infty) \times [u, \infty)$. In the bi-variate case, the cumulative distribution function (*c.d.f.*) H itself is easily representable and we consider the point-wise posterior predictive estimates $\hat{H}$.

The number of MCMC steps is set to the conservative value of 5×10^5 for the non-parametric model, and to 2×10^5 for both DM models. Figure 9 displays three examples of fit with one data set generated respectively from a logistic, an Asymmetric logistic and a DM distribution. The estimation errors $\hat{H} - H_m$ are plotted. In this bi-variate setting, with this large number of iterations, the two versions of the Dirichlet model produce very similar estimates, so that only the ones from the re-parametrized version are displayed and compared to the non-parametric counterpart. The possibility for point masses at the end points is an advantage in favor of the non-parametric model, when the underlying distribution presents such a feature (middle panel, Asymmetric logistic distribution). On the other-hand, when the true distribution is continuous on $[0, 1]$, this flexibility seems to become a drawback: the posterior estimate grants some mass to $\{0\}$ and $\{1\}$, whereas the true distribution does not.

For a more quantitative assessment, the performance of the posterior mean estimates $\hat{H}$ for a given 'true' H_m are compared in terms of mean integrated squared error loss (MISE), which is $\mathrm{MISE}\,(\hat{H}, H_m) = \int_0^1 [\hat{H}(w) - H_m(w)]^2 \,\mathrm{d}w$, and the

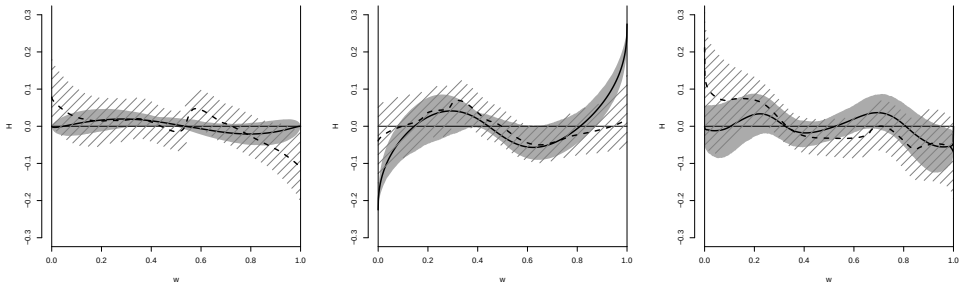

Figure 9: Error of the predictive angular *cdf* (solid lines) on the segment [0,1]. From left to right: data from a Logistic, an Asymmetric logistic and from a DM distribution. Solid line and gray area: Dirichlet Mixture mean estimate and $0.1 - 0.9$ posterior quantiles; dashed line and dashed area: *idem* in the non-parametric model.

scores are averaged over the 100 data sets, for each underlying distribution. Table 4 gathers the averaged MISE scores. For the sake of readability, the values have been multiplied by 10^3. As could be expected, the non-parametric estimator obtains the best score for the Asymmetric logistic model, because it allows point masses at the segment end-points. In the two other cases (no mass on the boundary), the converse is observed: the non-parametric estimate is outperformed by the DM model, probably for the same reason that makes the non-parametric framework preferable in the Asymmetric logistic case. As a conclusion for the bi-variate case, there is no clear general advantage in favor of one model against the others, and the original and re-parametrized versions of the DM model behave similarly, provided that the number of MCMC steps is large enough.

Table 4: Averaged MISE scores for the three inferential schemes (standard error of the estimate)

True distribution	Logistic	Asymmetric Logistic	Dirichlet Mixture
Re-parametrized DM	0.57 (0.05)	3.45 (0.18)	1.17 (0.1)
Original DM	0.63 (0.04)	3.58 (0.17)	0.96 (0.07)
Non-parametric	1.28 (0.07)	1.07 (0.08)	2.25 (0.17)

8. Discussion

In this paper, we demonstrate that Boldi and Davison (2007)'s model, can, after a suitable re-parametrization, be used in a Bayesian framework to infer the dependence structure between the largest observations of a multivariate data set of moderate dimension. For bi-variate problems, the DM model's performance is comparable to that of the fully non-parametric Bayesian model introduced by Guillotte et al. (2011): their relative goodness-of-fit scores depends on the true spectral distribution. The presence of point masses at the end points of the interval $[0, 1]$ induces a better fit of Guillotte et al. (2011)'s model, whereas the DM model obtains the best score when the true distribution is absolutely continuous. In this bi-variate setting, the original and the re-parametrized versions of the DM model produce very similar results, provided that the number of MCMC iterations is large enough. In greater dimension, the main added value of the re-parametrization is improved convergence of the reversible-jump algorithm, so that the generated Markov chains correctly span the support of the posterior and that estimated posterior credible sets are wider. Also, results for five-dimensional simulated data sets of size 100 indicate a rather low sensitivity of mean estimates to the specification of hyper-parameters: the mixing properties of the algorithm are enhanced if some prior dependence is introduced between the location for the mean vectors $\boldsymbol{\mu}$ and the weights, then the particular choice of hyper-parameters within a reasonable range does not significantly influence neither goodness-of-fit nor mixing properties.

The required computational effort is moderate; typical running times to issue the posterior samples on a desktop machine range from less than three minutes (for the three dimensional simulated data) to three hours (for the five dimensional Leeds data set). We have not tested the model on greater dimensional data sets, but much more than 100 data points would likely be needed to obtain reasonably precise results, and the computational time would naturally increase.

Supplementary material

An R package implementing the algorithm and the convergence assessment tools developed in this work has been prepared. It is available on demand to the authors and is intended to be submitted to the CRAN package repository.

Appendix A. Re-parametrization of the Dirichlet Mixture model

Expression for T_m. Recall that, from the definition, $\boldsymbol{\mu}_{\cdot,k} = \boldsymbol{\gamma}_{k-1}$, and that by (3), we have $\boldsymbol{\gamma}_0 = (1/d, \ldots, 1/d)$. Also, by associativity, for $1 \leq m \leq k-1$,

$$\rho_{m-1}\,\boldsymbol{\gamma}_{m-1} = p_m\,\boldsymbol{\mu}_{\cdot,m} + \rho_m\,\boldsymbol{\gamma}_m.$$

Both weights defining the center of mass $\boldsymbol{\gamma}_{m-1}$ are positive and, assuming (5), $\boldsymbol{\gamma}_{m-1}$ is on the line segment joining $\boldsymbol{\gamma}_m$ and $\boldsymbol{\mu}_{\cdot,m}$ (see Figure 1 for the three-dimensional case). Consequently,

$$\exists t_m > 0,\ \boldsymbol{\gamma}_m = \boldsymbol{\gamma}_{m-1} + t_m(\boldsymbol{\gamma}_{m-1} - \boldsymbol{\mu}_{\cdot,m}).$$

With the notations of section 3, $\mathcal{C}_m = \big\{i \in \{1,\ldots,d\} \ :\ \gamma_{i,m-1} - \mu_{i,m} < 0\big\}$. Thus, for $i \notin \mathcal{C}_m$, the map $t \mapsto \gamma_{i,m} + t(\gamma_{i,m} - \mu_{i,m})$ is non decreasing. Thus, $\forall i \notin \mathcal{C}_m, \forall t > 0, \gamma_{i,m} + t\,(\gamma_{i,m} - \mu_{i,m}) > 0$, whence

$$\begin{aligned}
T_m &= \sup\{t \geq 0 : \quad \forall i \in \mathcal{C}_m,\ \gamma_{i,m} + t\,(\gamma_{i,m} - \mu_{i,m}) > 0\} \\
&= \sup\left\{t \geq 0 : \quad t < \min_{i \in \mathcal{C}_m}\left(\frac{\gamma_{i,m}}{\mu_{i,m} - \gamma_{i,m}}\right)\right\} \\
&= \min_{i \in \mathcal{C}_m}\left(\frac{\gamma_{i,m}}{\mu_{i,m} - \gamma_{i,m}}\right).
\end{aligned}$$

Proof of Proposition 1. The equivalence of the two parametrizations is immediate from he argument preceding the proposition. Here, we derive the expression for p_m, given the current center of mass $\boldsymbol{\gamma}_{m-1}$, mean vector $\boldsymbol{\mu}_{\cdot,m}$ and eccentricity e_m, *i.e.* $p_m = \rho_m \frac{e_m T_m}{e_m T_m + 1}$.

Let h_θ a Dirichlet mixture density with parameter $\theta = \big(\boldsymbol{\mu}_{\cdot,1:k-1}, e_{1:k-1}, \nu_{1:k}\big) \in \boldsymbol{\Theta}_k$. Let $p_{1:k}$, $\boldsymbol{\mu}_{\cdot,k}$ be the corresponding weights vector and the "last" mean vector in the original parametrization . Let $m \geq 1$ and suppose the $p_j's$ $(j < m)$ have been reconstructed, so that $\rho_{m-1} = 1 - \sum_{j<m} p_j$. Since $\boldsymbol{\gamma}_{m-1} = \rho_{m-1}^{-1}\left\{p_m\,\boldsymbol{\mu}_{\cdot,m} + \rho_m\,\boldsymbol{\gamma}_m\right\}$, with $\rho_{m-1}^{-1}(p_m + \rho_m) = 1$, we have

$$\rho_{m-1}^{-1} p_m(\boldsymbol{\mu}_{\cdot,m} - \boldsymbol{\gamma}_{m-1}) + (1 - \rho_{m-1}^{-1} p_m)(\boldsymbol{\gamma}_m - \boldsymbol{\gamma}_{m-1}) = 0\ ,$$

whence

$$\rho_{m-1}^{-1} p_m(\boldsymbol{\mu}_{\cdot,m} - \boldsymbol{\gamma}_m) = \boldsymbol{\gamma}_{m-1} - \boldsymbol{\gamma}_m\ .$$

By assumption (5), $\boldsymbol{\mu}_{\cdot,m} \neq \boldsymbol{\gamma}_{m-1}$, so that $\boldsymbol{\gamma}_m \neq \boldsymbol{\gamma}_{m-1}$ and necessarily $\boldsymbol{\mu}_{\cdot,m} -$

$\boldsymbol{\gamma}_m \neq \mathbf{0}$. We thus have

$$\begin{aligned}
\rho_{m-1}^{-1} p_m &= \frac{\|\boldsymbol{\gamma}_m - \boldsymbol{\gamma}_{m-1}\|}{\|\boldsymbol{\gamma}_m - \boldsymbol{\mu}_{\cdot,m}\|} \\
&= \frac{e_m T_m \|\boldsymbol{\gamma}_{m-1} - \boldsymbol{\mu}_{\cdot,m}\|}{e_m T_m \|\boldsymbol{\gamma}_{m-1} - \boldsymbol{\mu}_{\cdot,m}\| + \|\boldsymbol{\gamma}_{m-1} - \boldsymbol{\mu}_{\cdot,m}\|} \\
&= \frac{e_m T_m}{e_m T_m + 1} \,.
\end{aligned}$$

□

Appendix B. Weak consistency of the posterior

The proof of Proposition 2 is an application of Schwartz's theorem (Schwartz, 1965, Theorem 6.1, *p.*22). The latter requires that the sample space $(S, \mathcal{S})$ be a separable, complete metric space, which is obviously the case with the simplex $\mathbf{S}_d$ endowed with the Euclidean metric and the Lebesgue σ-field . Let $\mathcal{M}$ be the set of absolutely continuous probability measures on S *w.r.t.* to some reference measure, which is in our case the Lebesgue measure on $\mathbf{S}_d = \{(w_1, \ldots, w_{d-1}) : w_i \geq 0, \sum_1^{d-1} w_i \leq 1\}$. A dominated statistical model is a subset $\mathcal{M}_\Theta = \{h_\theta, \theta \in \Theta\}$ of $\mathcal{M}$, indexed by some parameter space Θ. In a non parametric context, Θ is any measurable space with σ-field $\mathcal{T}$. The mapping $\theta \mapsto h_\theta$ defines a pre-image σ-algebra $\mathcal{T}'$ on $\mathcal{M}_\Theta$, so that a prior π on $\mathcal{T}$ induces a prior π' on $\mathcal{T}'$. For the sake of simplicity, we drop the $'$, so that $\mathcal{T}$ and π will respectively be used to denote the σ-field and the prior both on Θ and on $\mathcal{M}_\Theta$.

For us, $\Theta = \boldsymbol{\Theta}_B$ (defined in (10)) and $\mathcal{T}$ is the Borel σ-field associated with the topology induced by the Euclidean topology on the co-product space $\boldsymbol{\Theta}_B$.

In the following, it must be assumed that the function $(\mathbf{w}, \theta) \mapsto h_\theta(\mathbf{w})$ is $(\mathcal{S} \times \mathcal{T})$-measurable. This is the case when $\mathcal{M}_\Theta$ is the DM model. As for random variables, the infinite sequence $(\mathbf{W})_\infty = \{\mathbf{W}_j,\ j \geq 0\}$ corresponds to (*i.i.d.*) random vectors following the density $h_0 \in \mathcal{M}$ and $\mathbf{W}_{1:n} = (\mathbf{W}_1, \ldots, \mathbf{W}_n)$ to a sample of size n. Also, the same notation h_0 is used to refer to the distribution of $\mathbf{W}$, $\mathbf{W}_{1:n}$ or $\mathbf{W}_\infty$ (defined on the product σ-fields). Finally, π_n denotes the posterior $\pi(\cdot|\mathbf{W}_{1:n})$ on $\mathcal{T}$. The notion of *uniformly consistent sequence of tests* is key to establishing weak consistency. Consider the two sided hypothesis

$$\mathcal{H}_0 :\ h = h_0 \quad \text{versus} \quad \mathcal{H}_1 :\ h \in U^c \,,$$

where $U \subset \mathcal{M}$ and $h_0 \in U$. Let $(\tau_n)_{n \geq 1}$ be a sequence of tests (*i.e.*: τ_n is a function of $\mathbf{W}_{1:n}$), with $0 \leq \tau_n \leq 1$ aiming at testing $\mathcal{H}_0$ versus $\mathcal{H}_1$. Then, $(\tau_n)_n$ is said

uniformly consistent if

$$\mathbb{E}_{h_0}(\tau_n) \underset{n\to\infty}{\longrightarrow} 0, \text{ and } \inf_{h\in U^c} \mathbb{E}_h(\tau_n) \underset{n\to\infty}{\longrightarrow} 1.$$

Throughout her paper, Schwartz assumes that the model is identifiable. However, since we focus on weak consistency, we shall only need one of her results which we restate below for convenience and does not require identifiability. A self contained proof of this theorem may be found in Ghosh and Ramamoorthi (2003).

Theorem 1. *(L. Schwartz, 1965)*
Let π a prior *on $\mathcal{T}$ and $h_0 \in \mathcal{M}$. Let $U \subset \mathcal{M}$ containing h_0, such that $U \cap \mathcal{M}_\Theta$ be $\mathcal{T}$-measurable. If*

- *The application $(\mathbf{w}, \theta) \mapsto h_\theta(\mathbf{w})$ is $(\mathcal{S} \times \mathcal{T})$-measurable,*
- *h_0 is in the KL support of π,*
- *There is a uniformly consistent sequence of tests for*

$$\mathcal{H}_0 : \ h = h_0 \quad \textit{versus} \quad \mathcal{H}_1 : \ h \in \mathcal{M} \setminus U \ ,$$

Then

$$\pi_n(U \cap \mathcal{M}_\Theta) \underset{n\to\infty}{\longrightarrow} 1 \ , \ h_0\textit{-almost surely.} \tag{B.1}$$

The identifiability assumption is used in Schwartz's paper to exhibit a uniformly consistent sequence of test for metric neighborhoods. As we shall see, this is unnecessary for our purposes, because we consider only weak neighborhoods of the true density, so that uniformly consistent sequences of tests can always be constructed. Let $\mathcal{M}$ be endowed with the Borelian σ-field $\mathcal{B}(\mathcal{M})$ generated by the weak topology on $\mathcal{M}$. When $\Theta = \mathbf{\Theta}_B$ is the truncated parameter space for the DM model, it is easily verified that the intersections of open sets in $\mathcal{M}$ with $\mathcal{M}_\Theta$ are $\mathcal{T}$-measurable (if g is some bounded, continuous function on $\mathbf{S}_d$, the map $\theta \mapsto \int_{\mathbf{S}_d} g\, h_\theta$ is continuous on all compact subset of $\mathbf{\Theta}_B$). Consequently, a prior π on $(\mathbf{\Theta}_B, \mathcal{T})$ induces a prior $\tilde{\pi}$ on $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$ defined by $\tilde{\pi}(U) = \pi(U \cap \mathcal{M}_\Theta)$. Again, the $\tilde{}$ is omitted and π denote both the prior on $\mathcal{M}$ and on $\mathbf{\Theta}_B$. As noted *e.g.* in Ghosal et al. (1999), and shown in Ghosh and Ramamoorthi (2003), if U is a weak neighborhood of h_0 in $\mathcal{M}$, a uniformly consistent sequence of tests for $\mathcal{H}_0$ versus $\mathcal{H}_1$ is easily found. Indeed, any weak neighborhood may be obtained as a finite intersection of U's of the type $\{h \ : \ \int g\, h_0 - \int g\, h < \epsilon\}$, for some g bounded, continuous with $0 < g < 1$, so that, if τ_n is chosen as the indicator function of the set $\{\mathbf{W}_{1:n} \ : \ \frac{1}{n}\sum_1^n \mathbf{W}_i - \int h_0 g < \epsilon/2\}$, then $(\tau_n)_n$ is uniformly

consistent. Consequently, for such a U, the two first hypotheses in Theorem 1 imply the existence of a uniformly consistent sequence of tests, so that $\pi_n(U) \to 1$. For general weak neighborhoods $V = \bigcap_{r=1}^{R} U_r$, where U_r is as above, $\pi_n(V) \to 1$ as well.

Finally, since the sample space S is separable, the space of densities $\mathcal{M}$ is separable for the weak topology (see Billingsley, 1999, Theorem 6.8, for a proof that can easily be adapted to the case of absolutely continuous distributions). The weak neighborhoods of h_0 in $\mathcal{M}$ thus have a countable basis and we can exhibit a set $\Omega_0 \subset \mathcal{S}^{\mathbb{N}}$, with $h_0(\Omega_0) = 1$, on which convergence (B.1) occurs for *all* neighborhoods of h_0. We have shown (see also Ghosh and Ramamoorthi, 2003, chapter 4):

Corollary 1. *Let π be a prior on $(\Theta, \mathcal{T})$, with the regularity assumption:*
$\mathcal{B}(\mathcal{M}) \cap \{h_\theta, \theta \in \Theta\} \subset \mathcal{T}$ and $(\mathbf{w}, \theta) \mapsto h_\theta(\mathbf{w})$ is $(\mathcal{S} \times \mathcal{T})$-measurable.
If h_0 is in the KL support of π, then the posterior is weakly consistent at h_0.

Proposition 2 can now be proven.

Proof of Proposition 2. The regularity requirements for the corollary to apply are met. Thus, we only need to show that the KL closure of $\mathbf{\Theta}_B$ is included in the KL support of π. Let $h_0 \in \mathcal{M}$ be in the KL support of $\mathbf{\Theta}_B$. In other words, for any $\epsilon > 0$, we assume the existence of a $\theta_\epsilon \in \mathbf{\Theta}_B$ such that $KL(h_0, h_\theta) < \epsilon$.

Let $\epsilon > 0$ and $K_{h_0,\epsilon}$ a KL neighborhood of h_0: $K_{h_0,\epsilon} = \{h \in \mathcal{M} \ : \ KL(h_0, h) < \epsilon\}$. We need to show that $\pi(K_{h_0,\epsilon}) > 0$. By assumption (stated in the proposition), if U is a non empty open set in $\mathbf{\Theta}_B$, then $\pi(U) > 0$. Consequently, it is enough to exhibit a non empty open set $U^\epsilon \subset \mathbf{\Theta}_B$ (for the co-product Euclidean topology on Θ_B), such that $U^\epsilon \subset K_{h_0,\epsilon}$.

Let $k \leq k_{\max}$ such that $\theta_\epsilon \in \mathbf{\Theta}_k$. Then there is a closed ball $\bar{B}_\epsilon$ in $\mathbf{\Theta}_k$ (for the Euclidean metric), centered at θ_ϵ, such that $\bar{B}_\epsilon \subset \mathbf{\Theta}_k$. Let

$$\begin{aligned} \kappa : \bar{B}_\epsilon &\to \mathbb{R}^+ \\ \theta &\mapsto KL(h_0, h_\theta) \ . \end{aligned}$$

If we can show that κ is continuous on $\bar{B}_\epsilon$ for the Euclidean topology, then we are done. Indeed, continuity implies the existence a neighborhood $V^\epsilon \subset \bar{B}_\epsilon$ around θ_ϵ where $\kappa < \epsilon$, *i.e.* such that $V^\epsilon \in K_{h_0,\epsilon}$. Then one may choose $U^\epsilon = \mathbf{\Theta}_B \cap V^\epsilon$, where the intersection is non empty (clearly, $\mathbf{\Theta}_B$ has no isolated points in $\mathbf{\Theta}$). Let us now prove the continuity of κ. Let

$$
\begin{aligned}
g : \bar{B}_\epsilon \times \overset{\circ}{\mathbf{S}}_d &\longrightarrow \mathbb{R} \\
(\theta, \mathbf{w}) &\longmapsto \log\left(\frac{h_0(\mathbf{w})}{h_\theta(\mathbf{w})}\right) h_0(\mathbf{w}) \; ;
\end{aligned}
$$

so that $\kappa(\theta) = \int_{\overset{\circ}{\mathbf{S}}_d} g(\theta, \mathbf{w}) \, \mathrm{d}\mathbf{w}$. The function g is continuous in θ for all $\mathbf{w}$, and measurable in $\mathbf{w}$ for all θ. By continuity of the Lebesgue integral, we only need to show that g is uniformly dominated on $\bar{B}_\epsilon$ by some integrable function $g_0 : \overset{\circ}{\mathbf{S}}_d \to \mathbb{R}^+$. For such purpose, let us define

$$
\begin{aligned}
a_{\min} &= \min\left\{\mu_{i,m}\,\nu_m : \; m \leq k, i \leq d, (\,\boldsymbol{\mu}, \mathbf{e}, \boldsymbol{\nu}) \in \bar{B}_\epsilon\right\} > 0\,, \\
a_{\max} &= \max\left\{\mu_{i,m}\,\nu_m : \; m \leq k, i \leq d, (\,\boldsymbol{\mu}, \mathbf{e}, \boldsymbol{\nu}) \in \bar{B}_\epsilon\right\}, \\
D_{\min} &= \min\left\{\frac{\Gamma(\nu_m)}{\prod_{i=1}^d \Gamma(\mu_{i,m}\,\nu_m)} : m \leq k, i \leq d, (\,\boldsymbol{\mu}, \mathbf{e}, \boldsymbol{\nu}) \in \bar{B}_\epsilon\right\} > 0\,, \\
D_{\max} &= \max\left\{\frac{\Gamma(\nu_m)}{\prod_{i=1}^d \Gamma(\mu_{i,m}\,\nu_m)} : m \leq k, i \leq d, (\,\boldsymbol{\mu}, \mathbf{e}, \boldsymbol{\nu}) \in \bar{B}_\epsilon\right\} \; .
\end{aligned}
$$

Note that, by compactness of $\bar{B}_\epsilon$, the *extrema* are reached, which ensures positivity of the *infima*. Hence, $\forall(\,\boldsymbol{\mu}, \mathbf{e}, \boldsymbol{\nu}) \in \bar{B}_\epsilon, \forall \mathbf{w} \in \overset{\circ}{\mathbf{S}}_d$, $\forall m \leq k$,

$$
0 < D_{\min} \prod_{1 \leq i \leq d} w_i^{a_{\max}-1} \leq \mathrm{diri}(\mathbf{w} \mid \boldsymbol{\mu}_{\cdot,m}, \nu_m) \leq D_{\max} \prod_{1 \leq i \leq d} w_i^{a_{\min}-1} \; .
$$

By convex combination, we also have, $\forall \theta \in \bar{B}_\epsilon, \forall \mathbf{w} \in \overset{\circ}{\mathbf{S}}_d$,

$$
0 < D_{\min} \prod_i w_i^{a_{\max}-1} \leq h_\theta(\mathbf{w}) \leq D_{\max} \prod_i w_i^{a_{\min}-1} \; .
$$

Whence, by monotonicity of the log function, $\exists D_1, D_2 > 0$,

$$
D_1 + (a_{\max} - 1) \sum_i \log(w_i) \leq \log\left(h_\theta(\mathbf{w})\right) \leq D_2 + (a_{\min} - 1) \sum_i \log(w_i) \; .
$$

Let $C_1 = \max\left\{\,|D_1|\,, |D_2|\,\right\}$ and $C_2 = \max\left\{\,|a_{\min} - 1|\,, |a_{\max} - 1|\,\right\}$.
We have: $\forall(\theta, \mathbf{w}) \in \bar{B}_\epsilon \times \overset{\circ}{\mathbf{S}}_d$,

$$|\log(h_\theta(\mathbf{w})| \leq C_1 + C_2 \Big| \sum_{i=1}^{d} \log(w_i) \Big| \ .$$

Thus, $\forall(\theta, \mathbf{w}) \in \bar{B}_\epsilon \times \overset{\circ}{\mathbf{S}}_d$,

$$\begin{aligned} |g(\theta, \mathbf{w})| &\leq \left(|\log(h_0(\mathbf{w}))| + C_1 + C_2 \Big| \sum_{i=1}^{d} \log(w_i) \Big| \right) h_0(\mathbf{w}) \\ &= g_0(\mathbf{w}) \ . \end{aligned}$$

Using the fact that, for $\alpha > -1$, $w \mapsto w^\alpha \log(w)$ is integrable on $(0,1)$, with $w \mapsto \frac{1}{\alpha+1}(w^{\alpha+1}\log(s) - \frac{w^{\alpha+1}}{\alpha+1})$ as an anti derivative, g_0 is integrable on $\overset{\circ}{\mathbf{S}}_d$, so that κ is continuous on $\bar{B}_\epsilon$ and the proof is complete. □

Appendix C. Ergodicity properties of the Markov chain generated by the reversible jump algorithm.

In this section, $\tilde{\pi} = \pi_n$ denotes the posterior distribution and K is the M-DM kernel as defined in section 4 as a mixture kernel (one component corresponding to a given move choice).

Proof of Proposition 3.

Aperiodicity
It is enough to verify that, if $\theta_t \in \mathbf{\Theta}_B$, then the probability of rejecting the proposal is positive, *i.e.* $K(\theta_t, \{\theta_t\}) > 0$. This is true, *e.g.* because the probability of proposing a regular move is positive (and independent from θ_t) and the acceptance probability of a regular move is obviously strictly less than one.

η-irreducibility.
Here, the irreducibility measure is the Lebesgue measure on $\mathbf{\Theta}_B$, so that the prior π (hence, the posterior) and η are equivalent. In the sequel, let $\mathbf{\Theta}_{Bk}$ denote the index set of k-mixtures of Dirichlet densities in the prior's support: $\mathbf{\Theta}_{Bk} = \mathbf{\Theta}_B \cap \mathbf{\Theta}_k$. We need to show that, if $\theta_{\text{start}} \in \mathbf{\Theta}_{Bk}$ and $A \subset \mathbf{\Theta}_B$ is such that $\tilde{\pi}(A) > 0$, then there is a $i \geq 0$ such that $K^i(\theta_{\text{start}}, A) > 0$. The idea of the proof is very simple: we may choose A as a 'rectangular' subset of $\mathbf{\Theta}_{Bk'}$, for some $k' \leq k_{\max}$. If $k = k'$, we shall exhibit a finite sequence of regular move types (one move for each direction)

allowing to reach A from θ_{start}. If $k \neq k'$, it is easily verified that $\mathbf{\Theta}_{Bk'}$ is accessible from $\mathbf{\Theta}_{Bk}$. For the sake of completeness, we detail the proof.

For $\theta = (\boldsymbol{\mu}, e, \nu) \in \mathbf{\Theta}_{Bk}$, Let us organize the components of θ into $3k-2$ blocks $(\theta^1, \ldots, \theta^{3k-2})$, so that θ^m is respectively equal to $\boldsymbol{\mu}_{\cdot,m}$ (if $1 \leq m \leq k-1$), $e_{m-k_{\max}+1}$ (if $k \leq m \leq 2k-2$) or $\nu_{m-2k_{\max}+2}$ (if $2k-1 \leq m \leq 3k-2$). Similarly, we denote E_k^m the factor of the product space $\mathbf{\Theta}_{Bk}$ corresponding to direction m, so that $\mathbf{\Theta}_{Bk} = \prod_{m=1}^{3k-2} E_k^m$. Without loss of generality, take A as a 'rectangle' $A = \prod_{m=1}^{3k-2} A^m$, $A^m \subset E_k^m$.

Assume first that $\theta_{\text{start}} \in \mathbf{\Theta}_{Bk}$, and consider a sequence of move choices $c_{1:3k-2} = c_1, \ldots, c_{3k-2}$, made of all the possible regular move choices. The probability of such a sequence starting from θ_{start}, is non zero. If $x^m \in E_k^m$, let $\tilde{\theta}(\theta, x^m) = \left(\theta_1, \ldots, \theta^{m-1}, x^m, \theta^{m+1}, \ldots, \theta^{3k-2}\right)$ be the element of $\mathbf{\Theta}_{Bk}$ obtained by replacing some θ^m with x^m.

Finally, the probability of reaching A starting from θ_{start} is

$$K^{3k-2}(\theta, A|c_{1:3k-2}) \geq \int_{A_1} \cdots \int_{A_{3k-2}} \prod_{t=1}^{3k-2} q_t^k \, r_t^k \left(\theta_{t-1}, \tilde{\theta}(\theta_{t-1}, x^t)\right) \, \mathrm{d}x^1 \cdots \mathrm{d}x^{3k-2} \,,$$

where $\theta^0 = \theta_{\text{start}}$, and for $x^t \in E_k^t$, $\theta_t = \tilde{\theta}(\theta_{t-1}, x^t)$, q_t and r_t being the corresponding proposal density and acceptance probability, and, $\forall 1 \leq m \leq 3k-2$, $\theta^m \in A_1 \times \cdots \times A_m \times E_{m+1}^k \times \cdots \times E_{3k-2}^k$. Since each term of the product in the integrand is positive, we have $K^{3k-2}(\theta_{\text{start}}, A|c_{1:3k-2}) > 0$. Thus, $K^{3k-2}(\theta_{\text{start}}, A) > 0$.

Assume now that $\theta_{\text{start}} \notin \mathbf{\Theta}_{Bk}$. In such a case, the probability of proposing and accepting trans-dimensional moves until the chain reaches $\mathbf{\Theta}_{Bk}$ is positive. Consequently, $\mathbf{\Theta}_{Bk}$ is accessible from θ_{start}, which completes the proof.

Invariance of the posterior distribution under the M-DM kernel

Since the whole M-DM kernel K is a weighted average of partial kernels defined in section 4, it is enough to show that the posterior is invariant under each of them. The invariance under trans-dimensional moves is ensured by the fact that the acceptance ratios r_{split} and r_{combine} defined in section 4.2 satisfy Green (1995)'s balance condition. Also, each 'regular' kernel $K_m(\theta, \cdot)$ (*i.e.* affecting one $\boldsymbol{\mu}_{\cdot,m}$, one ν_m or one e_m corresponds to a *Metropolis-within-Gibbs* partial kernel, as defined *e.g.* in Roberts and Rosenthal (2006), so that, if we denote $\text{Im}(K_m, \theta) \subset \mathbf{\Theta}$ the image of $K_m(\theta, \cdot)$, η_m the reference Lebesgue measure on $\text{Im}(K_m, \theta)$, q_m the proposal density (w.r.t. η_m) and r_m acceptance probability, then, following Roberts and Rosenthal (2006, section 4), the so-called *balance equation*, $\tilde{\pi}(\theta) q_m(\theta, \theta^*) r_m(\theta, \theta^*) = \tilde{\pi}(\theta^*) q_m(\theta^*, \theta) r_m(\theta^*, \theta)$, ensures the invariance of $\tilde{\pi}$

under K_m.

We only need to show the invariance under the shuffle moves.
Let $r(\theta) = r_{\text{shuffle},m_1,m_2}(\theta, \theta^*)$ denote the acceptance probability of the *shuffle* move as described in section 4.3, for a transposition φ_{m_1,m_2}, so that $\theta^* = \varphi_{m_1,m_2}(\theta) := \varphi(\theta)$. Let K_s be the corresponding transition kernel (*i.e.*, the transition kernel conditionally to proposing a shuffle move affecting m_1 and m_2). We derive a sufficient condition on r for the posterior distribution π_n to be invariant under K_s. The proposal kernel Q_s, conditionally to the acceptance of the shuffle move, is the point mass $Q_s(\theta, A) = \delta_{\varphi(\theta)}(A) = \mathbf{1}_A(\varphi(\theta))$, for $A \subset \mathbf{\Theta}_B$. The shuffle kernel K_s may thus be written as

$$\begin{aligned} K_s(\theta, A) &= r(\theta)\mathbf{1}_A(\varphi(\theta)) + (1 - r(\theta))\mathbf{1}_A(\theta) \\ &= r(\theta)\mathbf{1}_{\varphi^{-1}(A)}(\theta) + (1 - r(\theta))\mathbf{1}_A(\theta) \ . \end{aligned}$$

and the shifted measure of A is

$$\begin{aligned} K_s.\pi_n(A) &= \int_{\varphi^{-1}(A)} \pi_n(\theta) r(\theta) \ \mathrm{d}\theta + \int_A (1 - r(\theta))\pi_n(\theta) \ \mathrm{d}\theta \\ &= \int_A \pi_n\left(\varphi^{-1}(\theta^*)\right) r(\varphi^{-1}(\theta^*)) \left|\mathrm{Jac}(\varphi)\right|^{-1}_{[\varphi^{-1}(\theta^*)]} \ \mathrm{d}\theta^* + \\ &\quad \ldots \int_A (1 - r(\theta))\pi_n(\theta) \ \mathrm{d}\theta \\ &= \pi_n(A) + \\ &\quad \ldots \int_A \pi_n\left(\varphi^{-1}(\theta^*)\right) r(\varphi^{-1}(\theta^*)) \left|\mathrm{Jac}(\varphi)\right|^{-1}_{[\varphi^{-1}(\theta^*)]} - r(\theta^*)\pi_n(\theta^*) \ \mathrm{d}\theta^* \ . \end{aligned}$$

A sufficient condition to have $K_s.\pi_n(A) = \pi_n(A)$ is thus that $\pi_n(\theta) r(\theta) \left|\mathrm{Jac}(\varphi)\right|^{-1}_{[\theta]} = r(\theta^*)\pi_n(\theta^*)$, or

$$\forall \theta \in \mathbf{\Theta}_B, \frac{r(\theta)}{r(\theta^*)} = \frac{\pi_n(\theta^*)}{\pi_n(\theta)} \left|\mathrm{Jac}(\varphi)\right|_{[\theta]} \tag{C.1}$$

Now, since φ is the transposition of two components of the Ψ-parametrization, we have $\varphi = \varphi^{-1}$, and

$$\left|\mathrm{Jac}(\varphi)\right|_{[\theta]} = \sqrt{\left|\mathrm{Jac}(\varphi)\right|_{[\theta]} \left|\mathrm{Jac}(\varphi)\right|_{[\theta]}} = \sqrt{\frac{\left|\mathrm{Jac}(\varphi)\right|_{[\theta]}}{\left|\mathrm{Jac}(\varphi^{-1})\right|_{[\varphi(\theta)]}}} = \sqrt{\frac{\left|\mathrm{Jac}(\varphi)\right|_{[\theta]}}{\left|\mathrm{Jac}(\varphi)\right|_{[\theta^*]}}} \ ,$$

so that (C.1) holds if we set $r(\theta)$ to

$$r(\theta) = \min\left(1, \frac{\pi_n(\theta^*)}{\pi_n(\theta)} \left|\mathrm{Jac}(\varphi)\right|_{[\theta]}\right)$$

Note that the above argument is not valid for general permutations of indices $\varphi_{m_1,\ldots,m_d}$, unless the condition $\varphi = \varphi^{-1}$ holds.

□

Appendix D. M-DM algorithm details

Appendix D.1. Proposal distribution for μ-moves

The proposal density $q_\mu(\boldsymbol{\mu}_{\cdot,m}(t), \cdot)$ is a Dirichlet mixture constructed from the data $\mathbf{W}_{1:n} = (\mathbf{W}_1, \ldots, \mathbf{W}_n)$:

$$q_\mu(\boldsymbol{\mu}_{\cdot,m}(t), \cdot) = \sum_{j=1}^{n} \tilde{p}_j \ \mathrm{diri}(\cdot \mid \tilde{\boldsymbol{\mu}}_{\mathbf{W}_j}, \tilde{\nu}).$$

The proposal parameters $(\tilde{\mathbf{p}}, \tilde{\boldsymbol{\mu}}_{\mathbf{W}}, \tilde{\nu})$ are as follows: Let $\tilde{\epsilon}_w$ be a recentring parameter, typically set to 0.1. Then

$$\tilde{\boldsymbol{\mu}}_{\mathbf{W}_j} = (1 - \tilde{\epsilon}_w)\mathbf{W}_j + \tilde{\epsilon}_w \boldsymbol{\gamma}_0 \,,$$

where $\boldsymbol{\gamma}_0 = (1/d, \ldots, 1/d)$. is the centroid of the simplex. The concentration parameter is set to $\tilde{\nu} = \frac{d}{\tilde{\epsilon}_w}$, So that each component $\mathrm{diri}(\cdot \mid \tilde{\boldsymbol{\mu}}_{\mathbf{W}_j}, \tilde{\nu})$ is bounded, with mode at $\mathbf{W}_j$. The weights $(\tilde{p}_1, \ldots, \tilde{p}_n)$ are defined so as to penalize the distance between $\boldsymbol{\mu}_{\cdot,m}(t)$ and $\mathbf{W}_j$. Namely, $\tilde{p}_j$ is proportional to the density, evaluated at $\mathbf{W}_j$, of a Dirichlet distribution with mode at $\boldsymbol{\mu}_{\cdot,m}(t)$. Again, we define $\tilde{\epsilon}_\mu \in (0, 1/2)$ (typically, $\tilde{\epsilon}_\mu = 0.1$), then $\tilde{\boldsymbol{\mu}}_\mu = (1 - \tilde{\epsilon}_\mu)\,\boldsymbol{\mu}_{\cdot,m}(t) + \tilde{\epsilon}_\mu \boldsymbol{\gamma}_0$ and $\nu^*_\mu = d/\epsilon^*_\mu$. Now, the un-normalized weight for the $j^{\ th}$ mean vector is

$$\tilde{\tilde{p}}_j = \mathrm{diri}(\mathbf{W}_j \mid \tilde{\boldsymbol{\mu}}_\mu, \tilde{\nu}_\mu).$$

Finally, we normalize the vector and set $\tilde{p}_j = \tilde{\tilde{p}}_j \,/\, \sum_{j=1}^{n} \tilde{\tilde{p}}_j$.

In short, the proposal mean vector $\boldsymbol{\mu}^*_{\cdot,m}$ has a good chance to be drawn in a small neighborhood of one data point $\mathbf{W}_j$, which in turn should be located close to the current mean vector $\boldsymbol{\mu}_{\cdot,m}(t)$.

Appendix D.2. Proposal distribution for split moves

The proposal distribution for the new mean vector $\boldsymbol{\mu}^*_{\cdot,k}$ is constructed similarly to the μ-moves distribution. Namely, the proposal density $q_{\mu,\mathrm{split}}$ is defined by

$$q_{\mu,\text{split}}(\theta_t,\,\cdot\,) = \sum_{j=1}^{n} \tilde{p}_j^{\text{split}}\ \text{diri}(\,\cdot\mid \tilde{\boldsymbol{\mu}}_{\mathbf{W}_j}, \tilde{\nu})$$

where the $\tilde{\boldsymbol{\mu}}_{\mathbf{W}_j}$'s and $\tilde{\nu}$'s are the same as in the μ-moves, and where the weights $\tilde{p}_j^{\text{split}}$ are defined in a similar way as the $\tilde{p}_j$'s, except that the recentring parameter $\tilde{\epsilon}_\mu = 0.1$ is replaced with $\tilde{\epsilon}_\mu^{\text{split}} = 0.5$ (except for the fit on Leeds data where we found that $\epsilon_\mu^{\text{split}} = 0.3$ was better) and that the 'current mean vector' $\boldsymbol{\mu}_{\cdot\,,\,m}(t)$ is replaced with the last vector $\boldsymbol{\mu}_{\cdot\,,\,k}(t)$ in the Ψ-parametrization. Compared to the μ-moves, the proposal distribution is thus less concentrated around $\boldsymbol{\mu}_{\cdot\,,\,k}(t)$. The kth eccentricity parameter e_k^* is generated, conditionally to the proposed mean vector $\boldsymbol{\mu}^*_{\cdot\,,\,k}$, according to the prior distribution:

$$q_{e,\text{split}}(\theta_t,\,\cdot\mid \boldsymbol{\mu}^*_{\cdot\,,\,k}) = \pi_{e,k}(\,\cdot\mid \boldsymbol{\mu}_{\cdot\,,\,1:k-1},\, \boldsymbol{\mu}^*_{\cdot\,,\,k},\, e_{1:k})\,.$$

Finally, the last shape parameter ν_{k+1}^* is generated according to the proposal distribution for regular ν-moves, conditionally on $\nu_k(t)$:

$$q_{\nu,\text{split}}(\theta_t,\,\cdot\,) = q_\nu(\nu_k(t),\,\cdot\,).$$

Appendix D.3. Jacobian term in the acceptance ratio for shuffle moves

Here is derived the closed form of $\text{Jac}(\varphi)$ appearing in (12). The indices m_1, m_2 are omitted, and we denote G the local diffeomorphism deduced from Γ:

$$\begin{aligned} G : \boldsymbol{\Theta}_{Bk} \subset \mathbb{R}^{3k-2} &\longrightarrow G(\boldsymbol{\Theta}_{Bk}) \subset \mathbb{R}^{3k-2} \\ \big(\boldsymbol{\mu}_{\cdot\,,\,1:k-1}, e_{1:k-1}, \nu_{1:k}\big) &\longmapsto \big(\boldsymbol{\mu}_{\cdot\,,\,1:k-1}, p_{1:k-1}, \nu_{1:k}\big)\ . \end{aligned}$$

Recall that $\varphi(\theta) = \Gamma^{-1} \circ \tau \circ \Gamma(\theta)$, where τ is the transposition of the directions corresponding to m_1 and m_2, so that

$$\text{Jac}(\varphi)_\theta = \text{Jac}(G^{-1})_{\tau\circ\Gamma(\theta)}\, \text{Jac}(\tau)_{\Gamma(\theta)}\, \text{Jac}(G)_\theta\,.$$

The determinant of τ is -1, so that

$$|\text{Jac}(\varphi)_\theta| = \left|\frac{\text{Jac}(G)_\theta}{\text{Jac}(G)_{\theta^*}}\right|\,,$$

and we only need to compute $\text{Jac}(G)$. The Jacobian matrix dG is of the form

$$dG = \begin{pmatrix} \mathbf{1}_{\mathbf{R}^{(d-1)(k-1)}} & 0 & 0 \\ M_{p,\mu} & M_{p,e} & 0 \\ 0 & 0 & \mathbf{1}_{\mathbf{R}^k} \end{pmatrix} ,$$

Where $\mathbf{1}_{\mathbf{R}^{(d-1)(k-1)}}$ denotes the identity matrix on $\mathbf{R}^{(d-1)(k-1)}$ and $M_{p,e}$ is the Jacobian matrix $\left(\frac{\partial p_i}{\partial e_j}\right)_{i,j<k}$ relative to $\mathbf{p}$ and $\mathbf{e}$. Hence, $\mathrm{Jac}(G) = |M_{p,e}|$. Since p_m depends only on the $\{\, \boldsymbol{\mu}_{\cdot,j}, e_j : j \leq m\}$, we have

$$|M_{p,e}| = \begin{vmatrix} \frac{\partial p_1}{\partial e_1} & 0 & \cdots & 0 \\ * & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \cdots & * & \frac{\partial p_{k-1}}{\partial e_{k-1}} \end{vmatrix} ,$$

whence

$$|M_{p,e}| = \prod_{m=1}^{k-1} \frac{\partial p_m}{\partial e_m} .$$

From Proposition 1, we have

$$\begin{aligned} \frac{\partial p_m}{\partial e_m} &= \frac{\partial}{\partial e_m} \left(\rho_{m-1} \frac{e_m T_m}{1 + e_m T_m} \right) \\ &= \frac{\rho_{m-1} T_m}{(1 + e_m T_m)^2} . \end{aligned}$$

Note that this holds because ρ_{m-1} and T_m do not depend on e_m: they are functions of the $\{\mu_j, e_j : \ j < m\}$ only.

The desired Jacobian's absolute value is thus

$$|\mathrm{Jac}(\varphi)| = \prod_{m=1}^{k-1} \frac{\rho_{m-1} T_m}{(1 + e_m T_m)^2} \prod_{m=1}^{k-1} \frac{(1 + e_m^* T_m^*)^2}{\rho_{m-1}^* T_m^*} , \tag{D.1}$$

where the $e_m^*, \rho_{m-1}^*, T_m^*$ are relative to the proposal parameter $\theta^* = \varphi(\theta)$.

Appendix E. Convergence assessment by integration against Dirichlet test functions

Appendix E.1. Random choice of Dirichlet test functions

This section details the procedure followed to construct a set of Dirichlet test functions $\{g_\ell = \mathrm{diri}(\,\cdot\,|\,\boldsymbol{\mu}_\ell, \nu_\ell),\ 1 \le \ell \le L\}$ that are used to monitor the chains convergence in our simulation study.

Let $\mathbf{W}_{1:n} = (\mathbf{W}_1, \ldots, \mathbf{W}_n)$ be an angular data set on which the model is to be fitted. In this study, we fix $L = 5$ and the $\tilde{\boldsymbol{\mu}}_\ell$'s are chosen so that they correspond to the dependence features of the data set (cf our remark preceding section 6.2). Namely, the $\tilde{\boldsymbol{\mu}}_\ell$'s are sampled among the angular data points as follows: A maximum shape parameter $\tilde{\nu}_{\max}$ is imposed, in order to exclude test functions inducing too large a variance for the empirical estimator $\hat{g}^{\mathrm{nonP}}$. In this study, we set $\nu_{\max} = 20\,d$, where d is the dimension of the sample space. The n' angular points $\mathbf{W}_j$ such that $\min_{1\le i\le d}\{W_{i,j}\} > 1/\tilde{\nu}_{\max}$ are retained as candidate data points, out of which L elements $(\mathbf{w}_{j_1}, \ldots, \mathbf{w}_{j_L})$ are drawn with equi-probability, and we set $\tilde{\boldsymbol{\mu}}_\ell = \mathbf{w}_{j_\ell}$. Finally, a minimum value $\tilde{\nu}_{\min} = 5 * d$ is imposed for the test's shape parameter (in order to avoid too flat test functions for points near the center of simplex), as well as a multiplying constant $\chi_{\mathrm{test}} = 1.001$, then the ℓth shape parameter is set to

$$\tilde{\nu}_\ell = \max\left\{\frac{\chi_{\mathrm{test}}}{\min_{1\le i\le d}\tilde{\mu}_{i,\ell}},\ \tilde{\nu}_{\min}\right\}.$$

Appendix E.2. Theoretical standard deviation of the empirical estimate of $\mathbb{E}_{h_0}(g)$, for g a Dirichlet test function.

Here, it is assumed that $h_0 = h_\theta$ is itself a Dirichlet mixture density. We already have the expression for $\mathbb{E}_\theta(g) = \mathbb{E}_{h_\theta}(g)$ when $g = \mathrm{diri}(\,\cdot\,|\,\tilde{\boldsymbol{\mu}}, \tilde{\nu})$ and $\theta = (\mathbf{p},\, \boldsymbol{\mu},\, \boldsymbol{\nu})$:

$$\mathbb{E}_\theta(g) = \sum_{m=1}^{k} p_m \mathcal{I}\,\boldsymbol{\mu}_{\cdot,m}, \nu_m(\,\tilde{\boldsymbol{\mu}}, \tilde{\nu})\ , \tag{E.1}$$

where the $\mathcal{I}\,\boldsymbol{\mu}_{\cdot,m}, \nu_m(\,\tilde{\boldsymbol{\mu}}, \tilde{\nu})$'s are given by (16). To compute $\mathbb{E}_\theta(g^2)$, we note that

$$g^2(\,\cdot\,) = C_{\tilde{\boldsymbol{\mu}},\tilde{\nu}}\,\mathrm{diri}(\,\cdot\,|\,\boldsymbol{\mu}', \nu')\ ,$$

with $\nu' = 2\tilde{\nu} - d$, $\boldsymbol{\mu}' = (2\tilde{\nu}\,\tilde{\boldsymbol{\mu}} - 1)/\nu'$ and $C_{\tilde{\boldsymbol{\mu}},\tilde{\nu}} = \frac{\Gamma(\tilde{\nu})^2}{\prod_{1\le i\le d}\Gamma(\tilde{\nu}\tilde{\mu}_i)^2}\frac{\prod_{1\le i\le d}\Gamma(\nu'\mu'_i)}{\Gamma(\nu')}$. The analytic expression for (18) follows:

$$\delta_n^{\mathrm{nonP}} = n^{-1/2} \left[C_{\tilde{\boldsymbol{\mu}},\tilde{\nu}} \sum_{m=1}^{k} p_m \mathcal{I}\, \boldsymbol{\mu}_{\cdot,m}, \nu_m(\boldsymbol{\mu}', \nu') - \left(\sum_{m=1}^{k} p_m \mathcal{I}\, \boldsymbol{\mu}_{\cdot,m}, \nu_m(\tilde{\boldsymbol{\mu}}, \tilde{\nu}) \right)^2 \right]^{1/2} .$$

Appendix F. Bi-variate distributions used in the simulation study

In this section, expressions for multivariate extreme value distributions are given for bi-variate vectors which uni-variate margins follow a unit-Fréchet distribution. For distributions of the logistic type, the angular density on $(0,1)$ is obtained using Theorem 1 from Coles and Tawn (1991) and the angular probability measure $H(w)$ follows by integration of the density between 0 and $w \in (0,1)$.

The first multivariate extreme value distribution used to generate data is a logistic one, with *cdf* of the type $F_{\mathrm{L}}(z_1, z_2) = \exp\left[-\left(z_1^{-1/r} + z_2^{-1/r}\right)^r\right]$, $(r \in (0,1]), z_i > 0$. If $r = 1$, the two variable are independent, lower values correspond to greater levels of dependence. For our simulation, we take $r = 0.6$.

The second distribution is an Asymmetric logistic one, characterized by the *cdf*

$$F_{\mathrm{AL}}(z_1, z_2) = \exp\left\{ -\frac{1-\theta_1}{z_1} - \frac{1-\theta_2}{z_2} - \left[\left(\frac{\theta_1}{z_1}\right)^{1/r} + \left(\frac{\theta_2}{z_2}\right)^{1/r} \right]^r \right\} .$$

The corresponding angular measure is

$$H_{\mathrm{L}}(w) = \frac{1}{2}\Big\{ 1 + \theta_1 - \theta_2 - \left[\theta_1^{1/r}(1-w)^{1/r-1} - \theta_2^{1/r} w^{1/r-1}\right] \times \cdots \\ \left[\theta_1^{1/r}(1-w)^{1/r} + \theta_2^{1/r} w^{1/r}\right]^{r-1} \Big\} .$$

The logistic distribution corresponds to the special case $\theta_1 = \theta_2 = 1$. Otherwise, the angular measure grants non-zero mass to the boundary points, $H_{\mathrm{AL}}(\{0\}) = (1-\theta_2)/2$ and $H_{\mathrm{AL}}(\{1\}) = (1+\theta_1)/2$. In this study, we set $r = 1/3$, $\theta_1 = 0.45$, $\theta_2 = 0.55$. For the logistic and the Asymmetric logistic distributions, data can easily be simulated using *e.g.* the `R` package `evd`. The marginal parameters are set in order to have unit-Fréchet margins, and the threshold (u, u) retained for fitting the models is the theoretical marginal 0.9 quantile, *i.e.* $u \simeq 9.49$. Conditionally on exceeding u, each marginal variable approximately follows a Generalized Pareto distribution (GPD): $P(X_j > x \,|\, X_j > u) = (1 + \xi \frac{x-u}{\sigma})^{-1/\xi}$ $(x > u, j \in \{1,2\})$, with

$\xi = 1$ and $\sigma = u$, so that the marginal parameters to be specified in Guillotte et al. (2011)'s model are $(\zeta_u = 0.1, \xi_j = 1, \sigma_j = u)$, where ζ_j is the marginal probability of an excess above u_j, ξ_j is the GPD shape parameter and σ_j is the scale parameter for the GPD above u_j. If n is the number of points belonging to the upper square $(u, \infty)^2$, then the Dirichlet mixture model is directly fitted on the angular data set $(W_{i,1}, W_{i,2}) = R_i^{-1}(X_{i,1}, X_{i,2})$,with $R_i = X_{i,1} + X_{i,2}$, $1 \leq i \leq n$, corresponding to the n points with largest radial component R.

The last angular distribution is a Dirichlet mixture H_{DM} with parameters

$$\boldsymbol{\mu} = \begin{pmatrix} 0.8 & 0.5 & 0.1 & 0.3 \\ 0.2 & 0.5 & 0.9 & 0.7 \end{pmatrix}, \ \mathbf{p} = (0.25, 0.5, 0.125, 0.125), \ \boldsymbol{\nu} = (20, 0.9, 1, 50).$$

Angular data points $\mathbf{W}_i = (W_{i,1}, W_{i,2})$ $(1 \leq i \leq 1000)$ are easily simulated from H_{DM}. To fit Guillotte et al. (2011)'s model, radial variables R_i $(1 \leq i \leq 1000)$ such that $P(R_i > r) = 1/r$ $(r > 1)$ are generated independently from the $\mathbf{W}_i$'s. The bi-variate points $\mathbf{Y}_i = R_i \mathbf{W}_i$ have marginal survival function $P(Y_{i,j} > y) = 1/(2y)$. Then, the marginal 0.9 quantile for $\mathbf{Y}_{i,j}$ is $\tilde{u} = 5$, and n is the number of $\mathbf{Y}_i$'s belonging to $(\tilde{u}, \infty)^2$. Again, the n angular points with largest radial component are retained to fit the Dirichlet mixture model. To fit Guillotte et al. (2011)'s model, the $\mathbf{Y}_{i,j}$'s exceeding $\tilde{u}$ are re-normalized (using probability integral transform) into Generalized Pareto variables $X_{i,j}$'s with arbitrary threshold $u = 10$, so that $P(X_{i,j} > x \,|\, X_{i,j} > u) = (1 + (x - u)/u)^{-1}$. The $Y_{i,j}$'s below $\tilde{u}$ are treated as left-censored data. Then, the marginal parameters for Guillotte et al. (2011)'s model are $(\zeta_u = 0.1, \xi_j = 1, \sigma_j = 10)$, $j \in \{1, 2\}$.

Appendix G. Comparison with the original Dirichlet mixture model

Appendix G.1. Erratum on the prior specification (Boldi and Davison, 2007)

In the original parametrization, the prior F_μ on μ is defined conditionally on the number k of mixture components and on the weights vector $\mathbf{p}$, by successive conditioning in the lexicographic order:

$$F_\mu(\boldsymbol{\mu}_{\cdot,1}, \ldots, \boldsymbol{\mu}_{\cdot,k} | k, \mathbf{p}) = f_{1,1}(\mu_{1,1})\, f_{1,2}(\mu_{1,2}|\mu_{1,1}) \cdots f_{1,k-1}(\mu_{1,k-1}|\mu_{1,1:k-2}) \cdots \\ f_{d-1,k-1}(\mu_{d-1,k-1}|\mu_{1:d-1,1:k-2}),$$

where $f_{i,j}$ is a uniform distribution on the largest interval $I_{i,j}$ $(i \leq d-1,\ j \leq k-1)$ allowing (3), and where the last column and the last line are deduced from the

others according to (3) and $\sum_i \mu_{i,m} = 1$. Boldi and Davison indicate zero as a lower bound for $I_{i,j}$. In fact, small values in the the first columns of $\boldsymbol{\mu}$ imply large ones on the last column, which, in some cases, induce negative values on the last line. It is left to the reader to verify that the correct lower bound for $I_{i,j}$ is

$$\max\left\{0, p_j^{-1}\left(d^{-1} - \sum_{m<j} p_m\,\mu_{i,m} - \sum_{m\in j+1,\ldots,k} p_m\,(1 - S_{i,m})\right)\right\}\ ,$$

where $S_{i,m} = \sum_{\ell<i} \mu_{\ell,m}$ $(1 \leq m \leq k)$.

Appendix G.2. Prior specification and MCMC *tuning parameters used in the simulations*

For comparison with the re-parametrized inferential scheme, the original version of the Bayesian model and the reversible-jump algorithm were re-implemented, following Boldi and Davison (2007) and Boldi (2004). For the sake of reproducibility, the numerical values for the hyper-parameters and the MCMC tuning parameters that were used in our simulations are gathered in this section.

The prior on the parameter $\psi = (k,\, \boldsymbol{\mu}, \mathbf{p},\, \boldsymbol{\nu})$ is of the form

$$\pi(\psi) = \pi_k(k)\,\pi_p(\mathbf{p}\,|\,k)\,F_\mu(\,\boldsymbol{\mu}\,|\,k, \mathbf{p})\,\pi_\nu(\,\boldsymbol{\nu}\,|\,k)\,.$$

π_k is a truncated Poisson distribution, with truncation bounds $(k_{\min}, k_{\max}) = (1, 15)$ and intensity $\lambda = 3$. π_p is the uniform distribution on the simplex $\mathbf{S}_k$, *i.e.* the Dirichlet distribution with parameter $\alpha = \nu\,\boldsymbol{\mu} = (1, \ldots, 1)$. F_μ is described in the preceding subsection, with the original error corrected. Finally, π_ν is a product of truncated log-normal distribution, with same bounds as in the re-parametrized version, $\log(\nu_m) \in (-2, \log(5000))$, Denoting (m_ν, σ_ν^2) the mean and variance for $\log(\nu_m)$, we set, following Boldi and Davison (2007) for the bi-variate case (section 7.5), $m_\nu = \log(2), \sigma_\nu = 50$. However, for higher dimensional data, we found that mixing properties and convergence were enhanced by setting these hyper-parameters to the same value as in the re-parametrized version, so that m_ν and σ_ν^2 are respectively set to $\log(10 * (d + 1))$ (where d is the dimension of the data) and $\log(1 + 5^2)$.

As for the MCMC scheme, we follow Boldi (2004), whose approach is summarized in Boldi and Davison (2007), Appendix B. Three types of moves are allowed, respectively called *split*, *combine* and *MCMC*. For the *split* and *MCMC* moves, three typical move sizes are allowed: *small*, *medium*, *big*. The *combine* moves are the simplest: a pair of mixture components (m_1, m_2) is randomly chosen, and the two corresponding mean vectors are combined into a single $\mu_{\cdot,m_0}$,

which is the center of mass for $((\mu_{\cdot,m_1}, p_{m_1}), (\mu_{\cdot,m_2}, p_{m_2}))$, with weight $p_{m_0} = p_{m_1} + p_{m_2}$. Then, $\log(\nu_{m_0})$ is drawn as a normal distribution with mean equal to $(\log(\nu_{m_1}) + \log(\nu_{m_2}))/2$, and variance set to $\log(1 + (s)^2)$, where s is respectively equal to $0.1, 0.3$ and 0.5 for a *small*, a *medium* or a *big* move. During a *split* move, one mixture component m_0 is split into two. For a *big* move, the proposal mean vector $\mu_{\cdot,m_2}$ is uniformly distributed on $\mathbf{S}_d$. For *small* (*resp. medium*) moves, $\mu_{\cdot,m_2}$ follows a Dirichlet distribution with mode at $\mu_{\cdot,m_0}$, and recentring parameter $\epsilon_\mu = 0.05$, (*resp.* $\epsilon_\mu = 0.3$), *i.e.* the mean vector for the proposal Dirichlet distribution is $\epsilon_\mu(1/d, \ldots, 1/d) + (1 - 1/\epsilon_\mu)\mu_{\cdot,m_2}$ and the concentration parameter is $\nu_\mu = d/\epsilon_\mu$. The weight p_{m_1} for the proposed component m_1 is determined by drawing $\mathbf{v} \in (0, 1)$ and letting $p_{m_1} = \mathbf{v}\, p_{m_0}$, then $p_{m_2} = p_{m_0} - p_{m_1}$. For a *big* move, ϵ_v is uniformly distributed. Otherwise, it follows a Beta distribution, with parameter $(a_1, a_2) = 2/\epsilon_v\, [\epsilon_v(1, 1) + (1 - \epsilon_v)(1, 0))]$, with ϵ_v respectively equal to 0.05 and 0.3 for a *small* (*resp. medium*) move. The position of $\mu_{\cdot,m_1}$ is defined so that the former mean vector $\mu_{\cdot,m_0}$ be the center of mass for the two proposals $\mu_{\cdot,m_1}$ and $\mu_{\cdot,m_2}$. Finally, the shape parameters ν_{m_1}, ν_{m_2} are proposed in a similar way as in the *combine* moves, with the mean of the log-transformed variables set to $\log(\nu_{m_0})$. During a *MCMC* move, a permutation $\{\sigma(1), \ldots, \sigma(k)\}$ of $\{1, \ldots, k\}$ is randomly chosen (by sampling without replacement in $\{1, \ldots, k\}$). Then, a *combine* move followed by a *split* move is successively applied to each pair $(\sigma(i), \sigma(j))$, for $i \in \{1, \ldots, k - 1\}$ and $j \in \{i + 1, \ldots, k\}$.

Acknowledgments

Part of this work has been supported by the EU-FP7 ACQWA Project (www.acqwa.ch), by the PEPER-GIS project, by the ANR (MOPERA, McSim, StaRMIP) and by the MIRACCLE-GICC project. The authors would like to thank Anne-Laure Fougères and Anthony Davison for their valuable advice, and Simon Guillotte for kindly providing the codes used in Guillotte et al. (2011)'s paper.

References

Asmussen, S. and Glynn, P. (2010). Harris recurrence and mcmc: A simplified approach. *Thiele Research Reports, Department of Mathematical Sciences, University of Aarhus.*

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004). *Statistics of extremes: Theory and applications.* John Wiley & Sons: New York.

Billingsley, P. (1999). *Convergence of probability measures.* Wiley Series in Probability and Statistics.

Boldi, M. (2004). *Mixture models for multivariate extremes.* PhD thesis, Ecole Polytechnique Federale de Lausanne.

Boldi, M.-O. and Davison, A. C. (2007). A mixture model for multivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):217–229.

Bunke, O. and Milhaud, X. (1998). Asymptotic behavior of Bayes estimates under possibly incorrect models. *The Annals of Statistics*, 26(2):617–644.

Coles, S. and Tawn, J. (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 377–392.

Cooley, D., Davis, R., and Naveau, P. (2010). The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis*, 101(9):2103–2117.

de Carvalho, M., Oumow, B., Segers, J., and Warchoł, M. (2013). A Euclidean likelihood estimator for bivariate tail dependence. *Communications in Statistics-Theory and Methods*, 42(7).

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory, An Introduction.* Springer Series in Operations Research and Financial Engineering.

Einmahl, J., de Haan, L., and Piterbarg, V. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *The Annals of Statistics*, 29(5):1401–1423.

Einmahl, J. and Segers, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics*, 37(5B):2953–2989.

Freedman, D. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386–1403.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Consistency issues in bayesian nonparametrics. In *Asymptotics, Nonparametrics and Time Series; A Tribute to Madan Lal Puri*, pages 639–667. Marcel Dekker.

Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian nonparametrics.* Springer.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Guillotte, S., Perron, F., and Segers, J. (2011). Non-parametric Bayesian inference on bivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).*

Heffernan, J. and Tawn, J. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, pages 1109–1144.

Meyn, S., Tweedie, R., and Glynn, P. (1993). *Markov chains and stochastic stability*. Springer London et al.

Resnick, S. (1987). *Extreme values, regular variation, and point processes, volume 4 of Applied Probability. A Series of the Applied Probability Trust*. Springer-Verlag, New York.

Resnick, S. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Series in Operations Research and Financial Engineering.

Roberts, G. and Rosenthal, J. (2004). General state space Markov chains and mcmc algorithms. *Probability Surveys*, 1:20–71.

Roberts, G. and Rosenthal, J. (2006). Harris recurrence of metropolis-within-gibbs and trans-dimensional Markov chains. *The Annals of Applied Probability*, 16(4):2123–2139.

Roberts, G. and Smith, A. (1994). Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. *Stochastic processes and their applications*, 49(2):207–216.

Rosenthal, J. (2001). A review of asymptotic convergence for general state space Markov chains. *Far East J. Theor. Stat*, 5:37–50.

Sabourin, A., Naveau, P., and Fougères, A.-L. (2013). Bayesian model averaging for multivariate extremes. *Extremes*, pages 1–26.

Schwartz, L. (1965). On Bayes procedures. *Probability Theory and Related Fields*, 4(1):10–26.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728.

Walker, S. (2004). Modern Bayesian asymptotics. *Statistical Science*, 19(1):111–117.

Chapitre 4

Estimation avec données historiques censurées

La confrontation aux « vraies données » peut réserver des surprises. Le cas d'étude hydrologique qui motive ce chapitre devait initialement avoir un rôle illustratif des méthodes développées plus haut. Il est apparu assez vite que le problème, proposé par Benjamin Renard, de l'estimation de la loi jointe des extrêmes de quatre séries temporelles de débits, avec présence de censures, nécessitait l'introduction d'outils statistiques supplémentaires.

Le problème des données censurées est récurrent en statistique, et les méthodes développées pour ce cas particulier ont une portée générale. Pour faciliter la diffusion des résultats aussi bien dans la communauté hydrologique que chez les statisticiens, il a été décidé de diviser le traitement de la question en deux articles. Le modèle statistique, l'algorithme et la vérification de la consistance du schéma d'augmentation de données ont été rédigés à part et un premier article (section 4.1) a été soumis à une revue de statistique. L'analyse détaillée des données des Gardons a été réalisée dans un second article (section 4.2) à soumettre très prochainement à une revue d'hyrologie.

4.1 Modèle censuré

SEMI-PARAMETRIC MODELLING OF EXCESSES ABOVE HIGH MULTIVARIATE THRESHOLDS WITH CENSORED DATA

ANNE SABOURIN

Abstract. One commonly encountered problem in statistical analysis of extreme events is that very few data are available for inference. This issue is all the more important in multivariate problems that the dependence structure among extremes has to be inferred. In some cases, *e.g.* in environmental applications, it is sometimes possible to increase the sample size by taking into account historical or incomplete series with partial censoring. In this work, a semi-parametric Dirichlet mixture model for multivariate extremes is adapted to the context of censored data and missing components. The censored likelihood, which is needed for Bayesian inference, has no analytic expression. A data augmentation scheme is introduced, which avoids multivariate integration of the Poisson process intensity over both the censored intervals and the failure region above threshold.

Multivariate extremes; censored data; semi parametric Bayesian inference; mixture models; Dirichlet mixture; reversible-jump algorithm.

1. Introduction

Data censoring is a recurrent issue in multivariate statistical analysis of extreme values. Observations may not be concomitantly extreme, so that the marginal Pareto model does not apply to all the margins forming a jointly extreme observation. A 'censored likelihood' approach (Smith, 1994; Ledford and Tawn, 1996; Smith et al., 1997) allows to take into account such partially extreme data: coordinates that do not exceed some large fixed threshold are simply considered as left-censored. Further, in many applications, the scarcity of extreme data induces a large amount of uncertainty in the estimation of the marginal parameters and the dependence structure of extremes. To obtain larger sample sizes, one option is to take into account longer series, but this results in a certain number of censored and missing data. What motivated this work is a hydrological application: The data set consists of water discharge at four neighbouring stations in the region of the Gardons, in the south of France. Daily maxima have been systematically recorded on the recent period (20th c.) and historical information is also available: unusually high water levels occurring before 1891 (the earliest one is dated from 1604) have been recorded and appear in the archives. A large part of the historical data is censored: only major floods are recorded, sometimes as an interval (*e.g.* 'the water level exceeded the parapet but the Mr. X's house was spared'). These events are followed by long periods during which nothing is known except that the water level did not exceed that of the previous flood. Uni-variate analysis for this data set has been carried on by Neppel et al. (2010). The aim of the present paper is to propose a new methodology for inferring the multivariate structure of

Date: June 14, 2013.

extremes using such censored data sets. The inferential framework is tested on simulated data; a detailed analysis of the hydrological data is reserved for future study.

One reason for estimating the dependence structure of extremes is that it provides probabilities of concomitant excesses of high thresholds by several variables, such as *e.g.* the water level at different sites in a region of interest. Another reason is that neglecting the dependence between neighbouring sites may alter the estimation of the marginal distributions of excesses, if one assumes that some parameters characterising the margins are constant over a coherent region. This assumption is the starting point of the so-called 'regional frequency analysis' in hydrology, which intent is to use all the data recorded over a region of interest to estimate some shared parameters. Our aim is to allow the combination of the two approaches (regional and historical data analysis) by modelling altogether the marginal parameters and the dependence structure of multivariate peaks over thresholds when the data are partially censored.

Under a standard assumption of multivariate regular variation (see Section 2), the dependence structure of extremes is entirely determined by an *angular measure* H, which is a finite measure on the unit sphere, and only has to satisfy a first moments constraint. To wit, H is the distribution of the directional component of re-scaled observations above high thresholds.

For applied purposes, it is common practice to use a parametric model of multivariate extremes. A widely used one is the Logistic model and its asymmetric and nested extensions (Gumbel, 1960; Coles and Tawn, 1991; Stephenson, 2009, 2003). In the logistic family, the dependence is represented *via* the exponent function, which is an integral form of the angular measure. The main advantage is that the censored versions of the likelihood are readily available, but parameters are subject to non linear constraints and structural modelling choices have to be made *a priori*, *e.g.*, by allowing only bi-variate or tri-variate dependence between closest neighbours.

However, the family of admissible dependence structures is, by nature, too large to be fully described by any parametric model. This pleads in favour of non parametric inference (Einmahl et al., 2001; Einmahl and Segers, 2009; Guillotte et al., 2011), or for semi-parametric compromises with the use of mixture models, built from a potentially infinite number of parametric components, such as the Dirichlet mixture model (DM), first introduced by Boldi and Davison (2007). They have shown that it can approach arbitrarily well any valid angular measure for extremes. A re-parametrised version of the DM model (Sabourin and Naveau, 2013) allows for Bayesian inference with a varying number of mixture components, with data sets of moderate dimension (typically, $d \approx 5$).

When censoring occurs, the likelihood expression involves the integral of the limiting density over rectangular regions. This constitutes a major impediment to the use of non parametric Bayesian models constructed *via* the angular measure. In this paper, the Bayesian DM model is adapted to the case of censored and missing data. The MCMC algorithm described in Sabourin and Naveau (2013) is modified to allow simultaneous inference of the dependence and marginal parameters. Instead of rectangular integration, a data augmentation scheme is used. Thus, the parameter space is embedded into a larger one, on which MCMC sampling is possible, and the Markov chain is built so that the posterior distribution on

the original parameter space is obtained by marginalisation of the shift-invariant measure defined on the augmented space.

The rest of this paper is organised as follows: Section 2 recalls the necessary probabilistic background for extreme values modeling. A Poisson model for excesses is introduced, the main features of the Dirichlet mixture model and its re-parametrised version are sketched and the Poisson joint likelihood for regular, non censored data is written. Censoring is introduced in Section 3, leading to a censored likelihood without analytic expression; The general principles of data augmentation are recalled and a data augmentation scheme is proposed for the problem under study. In Section 4, a reversible-jump MCMC algorithm allowing Bayesian inference in moderate dimension is described. The method is illustrated by a simulation study in Section 5 and Section 6 concludes.

2. Model for threshold excesses

2.1. **Notations and probabilistic framework.** In this paper, the sample space is the d-dimensional Euclidean space $\mathbb{R}^d$, endowed with the Borel σ-field. Let $(\mathbf{Y}_t)_{t\in\mathbb{N}}$ be independent, identically distributed (*i.i.d.*) random vectors in $\mathbb{R}^d$. To avoid technicalities, each $Y_{j,t}$ ($j \in \{1,\dots d\}$) is assumed non negative. In what follows, bold symbols denote vectors and $\mathbf{0}$ stands for $\mathbf{0}_{\mathbb{R}^d}$. The appropriate space to deal with convergence of extremes is $\mathbf{E} = [0,\infty]^d \setminus \{\mathbf{0}\}$. Weak convergence is denoted by $\overset{w}{\to}$. The space $\mathcal{M}^+(\mathbf{E})$ of non negative Radon measures on $\mathbf{E}$ is endowed with the topology induced by vague convergence[1], which is denoted by $\overset{v}{\to}$. Thus $\mathcal{M}^+(\mathbf{E})$ is a Polish space and can be endowed with the Borel σ-field $\mathscr{M}^+(\mathbf{E})$ corresponding to the vague topology. In this context, letting $(\Omega, \mathcal{A}, \mathbf{P})$ be the underlying probability space, a random measure is a $(\mathcal{A}, \mathscr{M}^+(\mathbf{E}))$- measurable map from Ω to $\mathcal{M}^+(\mathbf{E})$. Unless otherwise mentioned, $N(A)$ will denote the number of points observed in a region A.

The minimal requirement for extreme value theory to apply is that the $\mathbf{Y}_t$'s be in the domain of attraction (DOA) of a multivariate max-stable distribution G, *i.e.* that there exist $\mathbb{R}^d$ valued scaling sequences $(\mathbf{a}_n) = (a_{1,n},\dots,a_{d,n})_n, (\mathbf{b}_n) = (b_{1,n},\dots,b_{d,n})_n$, with $a_{j,n} > 0$, and a random vector $\mathbf{Y}_\infty$, with non degenerate cumulative distribution function G, such that

$$\bigvee_{t=1}^{n} \frac{\mathbf{Y}_t - \mathbf{b}_n}{\mathbf{a}_n} \underset{n\to\infty}{\overset{w}{\longrightarrow}} \mathbf{Y}_\infty \,. \tag{2.1}$$

Here, weak convergence occurs in $\mathbf{E}$ and $\bigvee$ denotes the component-wise maximum operator.

From Resnick (1987), Proposition 5.10, (2.1) is equivalent to the marginal distributions F_j belonging to the DOA of a uni-variate max-stable distribution, and the Fréchet-transformed data

$$\mathbf{X}_t = (-1/\log(F_1(Y_{1,t})),\dots,-1/\log(F_d(Y_{d,t}))) \tag{2.2}$$

[1]If $(\lambda_n)_n$ and λ are elements of $\mathcal{M}^+(\mathbf{E})$, vague convergence of λ_n to λ means that, for all relatively compact $B \subset \mathbf{E}$ such that $\lambda(\partial B) = 0$, $\lambda_n(B) \to \lambda(B)$.

being in the DOA of a multivariate max-stable law with unit-Fréchet margins. The latter condition is equivalent to weak convergence of the point process of rescaled Fréchet variables towards a Poisson random measure (PRM) on $[0,1] \times \mathbf{E}$. More precisely, (2.1) and marginal convergence are equivalent to (see *e.g.* Resnick, 1987, 2007; Coles and Tawn, 1991):

$$\sum_{t=1}^{n} \mathbb{1}_{(\frac{t}{n}, \frac{\mathbf{X}_t}{n})} \xrightarrow{w} \mathrm{PRM}(\ell \otimes \lambda) \tag{2.3}$$

in $\mathcal{M}^+([0,1] \times \mathbf{E})$, where ℓ denotes the Lebesgue measure on $\mathbb{R}$. The so-called *exponent measure* λ is homogeneous of order -1, which is conveniently expressed in polar coordinates: let $\|\cdot\|$ be any norm on $\mathbb{R}^d$, and $\mathbf{S}_d$ be the positive orthant of the unit sphere for this norm. Then, the image measure of λ by the polar transformation $\mathbf{x} \mapsto (r, \mathbf{w}) = (\|\mathbf{x}\|, \frac{\mathbf{x}}{\|\mathbf{x}\|}) \in (0, \infty] \times \mathbf{S}_d$, is a product measure. Namely, identifying the image measure with λ for the sake of conciseness, one may write

$$\mathrm{d}\lambda(r, \mathbf{w}) = \frac{c}{r^2}\ \mathrm{d}r\, \mathrm{d}H(\mathbf{w})\,, \tag{2.4}$$

where H is the angular measure and c is a positive normalising constant such that H be a probability measure. Concentration of mass in the middle of the unit sphere's positive orthant indicates strong dependence at extreme levels, whereas Dirac masses on the vertices characterises asymptotic independence. This paper focuses on the case where H grants some mass to the interior of positive orthant, so that all the variables are asymptotically dependent. Choosing the L_1 norm $\|\mathbf{x}\|_1 = \sum_{i=1}^{d} |x_j|$, the unit sphere is the unit simplex: $\mathbf{S}_d = \{(w_1, \ldots, w_d) : w_j \geq 0\,,\ \sum_{j=1}^{d} w_j = 1\}$, and $c = d$.

Due to the standard form of the $\mathbf{X}_j$'s, and to our choice of the L_1 norm, a probability measure H on $\mathbf{S}_d$ is a valid angular measure if and only if

$$\int_{\mathbf{S}_d} w_j\ \mathrm{d}H(\mathbf{w}) = \frac{1}{d} \quad (1 \leq j \leq d)\,. \tag{2.5}$$

Note that (2.5) is the only constraint on H, so that the angular measure has no reason to be part of any particular parametric family.

In the remainder of this paper, angular measure densities will be given with respect to the $d-1$ dimensional Lebesgue measure $\mathrm{d}\mathbf{w} = \mathrm{d}w_1 \cdots \mathrm{d}w_{d-1}$.

2.2. **Statistical modeling.** The following model for threshold excesses borrows both from the point process approach (see *e.g.* Coles and Tawn, 1991) and from the censored likelihood one (as *e.g.* in Ledford and Tawn, 1996). Considering (2.3), one may define a high multivariate threshold $\mathbf{v} = (v_1, \ldots, v_d)$ and, if n is the number of observed data, consider the failure region $A_{\mathbf{v}} = \mathbf{E} \setminus [\mathbf{0}, \mathbf{v}]$, where $[\mathbf{0}, \mathbf{v}] = [0, v_1] \times \cdots \times [0, v_d] \setminus \{\mathbf{0}\}$. We call 'excess above $\mathbf{v}$' any point in $A_{\mathbf{v}}$, and 'marginal excess' any $Y_{j,t} > v_j$. Marginal excesses are assumed to be independent and identically distributed (*i.i.d*) according to a generalised Pareto distribution with shape ξ_j and scale σ_j. In the hydrological context of regional frequency analysis, coherent regions are defined so that the shape parameter ξ be chosen constant over the considered sites. Thus, in the following, we assume $\xi_1 = \cdots = \xi_d = \xi$. Let $F_j^{\mathbf{v}}$ denote the jth marginal distribution conditionally on Y_j not exceeding v_j. The

marginal probability of exceeding the threshold is $\zeta_j = \mathbf{P}(Y_j > v_j)$ $(1 \le j \le d)$. The marginal parameters are gathered into a $(d+1)$-dimensional vector

$$\chi = (\log(\sigma_1), \ldots, \log(\sigma_d), \xi) \in \mathbb{R}^{d+1}.$$

The marginal model is thus

$$\begin{aligned} F_j^{(\chi)}(y) &= \mathbf{P}(Y_{j,t} \le y \,|\, \xi, \sigma_j) \\ &= \begin{cases} 1 - \zeta_j \left(1 + \xi \frac{y - v_j}{\sigma_j}\right)^{-1/\xi} & (y \ge v_j), \\ (1 - \zeta_j) F_j^{\mathbf{v}}(y) & (y < v_j). \end{cases} \end{aligned} \tag{2.6}$$

It is common practice (Coles and Tawn, 1991; Davison and Smith, 1990) to use an empirical estimate $\boldsymbol{\zeta}$ for the vector of probabilities of an excess, and to ignore any estimation error. The Fréchet-rescaled multivariate threshold is

$$\boldsymbol{u} = \mathbf{T}(\mathbf{v}) = -1/\log(1 - \boldsymbol{\zeta})$$

and does not depend on χ. Applying the marginal transformations

$$\mathcal{T}_j^{\chi}(y) = -1/\log\left(F_j^{(\chi)}(y)\right),$$

the marginal variables $X_{j,t} = \mathcal{T}_j^{\chi}(Y_{j,t})$ follow unit Fréchet distributions, $\mathbf{P}(X_{j,t} \le x) = \exp(-\frac{1}{x})$. The marginal transformations above threshold have inverse Jacobian

$$J_j^{\chi}(y_j) = \sigma_j^{-1} (\zeta_j)^{-\xi} x_j^2 e^{\frac{1}{x_j}} \left[1 - e^{\frac{-1}{x_j}}\right]^{1+\xi},$$

which will appear in the likelihood contribution of marginal excesses. In the sequel, $\mathbf{T}^{\chi}$ denotes the vector transformation: $\mathbf{T}^{\chi}(\mathbf{y}) = \left(\mathcal{T}_1^{\chi}(y_1), \ldots, \mathcal{T}_d^{\chi}(y_d)\right)$.

Let us introduce, on the Fréchet scale, the region

$$A_{\boldsymbol{u},n} = \frac{1}{n} \mathbf{T}^{\chi}(A_{\mathbf{v}}) = [0, \infty]^d \setminus [0, \frac{u_1}{n}] \times \cdots \times [0, \frac{u_d}{n}],$$

and $A_{\boldsymbol{u}} = A_{\boldsymbol{u},1}$. The $(\frac{t}{n}, \frac{\mathbf{X}_t}{n})$'s such that $\mathbf{X}_t \in A_{\boldsymbol{u}}$ are assumed to be the points of a Poisson process over $[0,1] \times A_{\boldsymbol{u},n}$, with intensity measure as in (2.3). The Poisson model is thus

$$\sum_{t=1}^{n} \mathbb{1}_{(\frac{t}{n}, \frac{\mathbf{X}_t}{n})}(\,\cdot\,) \sim \mathrm{PRM}(\ell \otimes \lambda) \text{ on } [0,1] \times A_{\boldsymbol{u},n}, \tag{2.7}$$

where λ is of the form (2.4).

2.3. **Dirichlet mixture angular measures.** In this paper, the angular measure H is modelled by a Dirichlet mixture distribution (Boldi and Davison, 2007; Sabourin and Naveau, 2013).

A Dirichlet distribution can be characterised by a shape $\nu \in \mathbb{R}^+$ and a centre of mass $\boldsymbol{\mu} \in \mathbf{S}_d$, so that its density is , for $\mathbf{w} \in \mathbf{S}_d$,

$$\mathrm{diri}_{\nu,\boldsymbol{\mu}}(\mathbf{w}) = \frac{\Gamma(\nu)}{\prod_{j=1}^{d} \Gamma(\nu \mu_j)} \prod_{j=1}^{d} w_j^{\nu \mu_j - 1}. \tag{2.8}$$

A parameter for a k-mixture is of the form

$$\psi = \big((p_1, \ldots, p_k), (\boldsymbol{\mu}_{\cdot,1}, \ldots, \boldsymbol{\mu}_{\cdot,k}), (\nu_1, \ldots, \nu_k)\big) \ ,$$

with weights $p_m > 0$, such that $\sum_m p_m = 1$. This is summarised by writing $\psi = \big(p_{1:k}, \boldsymbol{\mu}_{\cdot,1:k}, \nu_{1:k}\big)$. The corresponding mixture density is

$$h_\psi(\mathbf{w}) = \sum_{m=1}^{k} p_m \, \mathrm{diri}_{\nu, \boldsymbol{\mu}_{\cdot,m}}(\mathbf{w}) \ . \tag{2.9}$$

Condition (2.5) is satisfied if and only if

$$\sum_{m=1}^{k} p_m \boldsymbol{\mu}_{\cdot,m} = (1/d, \ldots, 1/d) \ . \tag{2.10}$$

Which, in geometric terms, means that the centre of mass of the $\boldsymbol{\mu}_{\cdot,1:m}$'s, with weights $p_{1:m}$, must lie at the centre of the simplex. As mentioned in the introduction, the family of Dirichlet mixture densities satisfying (2.10) is weakly dense in the space of admissible angular measure, as established by Boldi and Davison (2007). In addition, in a Bayesian framework, Sabourin and Naveau (2013) have shown that the posterior is weakly consistent under mild conditions. These two features put together make the Dirichlet mixture model an adequate candidate for modeling the angular components of extremes.

2.4. **Joint likelihood of uncensored data.** Let $\theta = (\chi, \psi)$ be the parameter for the joint model. Since marginal transformations are applied in the Cartesian coordinate system, it is convenient to express the Poisson likelihood in Cartesian coordinates as well. In addition, censoring will involve integration along the Cartesian axes. Thus, one needs the expression for the density of the exponent measure λ with respect to the d- dimensional Lebesgue measure $\mathrm{d}\mathbf{x} = \mathrm{d}x_1 \cdots \mathrm{d}x_d$, that is (Coles and Tawn, 1991, Theorem 1),

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\mathbf{x}}(\mathbf{x}) = d \, . \, r^{-(d+1)} h(\mathbf{w}) \, .$$

In particular, if $h = h_\psi$ is a Dirichlet mixture density as in (2.9), then

$$\frac{\mathrm{d}\lambda_\psi}{\mathrm{d}\mathbf{x}}(\mathbf{x}) = d \sum_{m=1}^{k} \left\{ \frac{p_m \Gamma(\nu_m)}{\prod_{j=1}^{d} \Gamma(\nu_m \mu_{j,m})} \prod_{j=1}^{d} x_j^{\nu_m \mu_{j,m} - 1} \left(\sum_{j=1}^{d} x_j \right)^{-(\nu_m + 1)} \right\} . \tag{2.11}$$

Let $t_1, \ldots, t_{n_\mathbf{v}}$ be the times of occurrence of an excess, *i.e.* $\mathbf{Y}_{t_i} \in A_\mathbf{v}$. In the simplified case where the $\mathbf{Y}_{j,t}$'s are all observed and where the marginals F_j's below threshold are known, the likelihood of the Poisson process is

$$\mathcal{L}_\mathbf{v}\left(\{\mathbf{y}_t\}_{1 \le t \le n}, \theta\right) \propto e^{-n \, \lambda_\psi(A_\boldsymbol{u})} \prod_{i=1}^{n_\mathbf{v}} \left\{ \frac{\mathrm{d}\lambda_\psi}{\mathrm{d}\mathbf{x}}(\mathbf{x}_{t_i}) \prod_{j : y_{j,t_i} > v_j} J_j^\chi(y_{j,t_i}) \right\}, \tag{2.12}$$

where the exponential term follows from

$$e^{-\ell([0,1]) \, . \, \lambda_\psi(A_{\boldsymbol{u},n})} = e^{-n \lambda_\psi(A_\boldsymbol{u})} \, .$$

3. Censored model

Two types of censoring are present: first, data are partially observed, which results in interval- or right-censoring. In addition, observed data points that exceed at least one threshold in one direction do not necessarily exceed all thresholds, so that the marginal extreme value model does not apply. Following Ledford and Tawn (1996), those components are also considered as left-censored.

The resulting censoring process $\mathscr{C}$ is assumed to be non informative. This means that (see also Gómez et al., 2004) that, if F is the marginal *c.d.f.* for Y_j and f is the marginal density, then Y_j's distribution conditional on having observed only the left and right censoring bounds (L, R) is $f/[F(R) - F(L)]$. This definition is easily extended to the multivariate case by replacing $F(R) - F(L)$ by the integral of the density over the censored directions (see *e.g.* Schnedler, 2005, for a proof of consistency of maximum censored likelihood estimators).

3.1. **Observed data.** This section accounts for 'natural censoring' which occurs independently of the choice of an extreme threshold $\mathbf{v}$ by the statistician. The observed process is denoted $\mathbf{O} = (\mathbf{O}_t)_t$, with $\mathbf{O}_t = (O_{1,t}, \ldots, O_{d,t})$. Marginal data $O_{j,t}$ $(1 \leq j \leq d)$ are classified into four types. Type 0 indicates a missing marginal record, type 1 is that of 'exact' (uncensored) data, type 2 and 3 are reserved respectively to right censored data, (when $Y_{j,t}$ is known to be greater than some lower bound L_j), and to doubly censored data, (when $Y_{j,t}$ is known to be comprised between a lower (possibly 0) and an upper bound: $Y_{j,t} \in [L_j, R_j]$).

Omitting the temporal index t, a marginal observation in direction j $(1 \leq j \leq d)$ is thus a 4-uple $O_j = (\kappa_j, Y_j, L_j, R_j)$, where κ, Y, L and R stand respectively for the data type, the recorded value (or some arbitrary value if $\kappa \neq 1$, denoted $\mathtt{NA}$), the lower bound (set to 0 if missing), and the upper bound (set to $+\infty$ if missing).

In this context, the definition of the position of a marginal data relatively to a given threshold v_j requires some care. The different situations are summarised in Figure 1. O_j exceeds v_j (left panel) if $\kappa_j = 1$ and $Y_j > v_j$, or if $\kappa_j \in \{2, 3\}$ and $L_j > v_j$. Similarly, O_j is below v_j if $\kappa_j = 1$ and $Y_j \leq v_j$, or if $\kappa_j = 3$ and $R_j \leq v_j$. If none of the above conditions hold, *i.e.* when censoring occurs with censoring interval intersecting the threshold, the relative positions of Y_j and v_j are unknown, and we say that O_j has *undetermined position* with respect to v_j.

For a multivariate observation $\mathbf{O}_t$, if at least one coordinate is undetermined or missing, and if the others are below the threshold, then $\mathbf{O}_t$ has undetermined position with respect to $\mathbf{v}$.

3.2. **Censoring observations below threshold.** Since the marginal distributions $F_j^{\mathbf{v}}$'s, conditional upon not exceeding v_j, are unknown, the $X_{j,t}$'s such that $Y_{j,t} < v_j$ are not available. Instead of attempting to estimate the $F_j^{\mathbf{v}}$'s, one can choose to censor the components below threshold. Namely, for a raw observation $O_{j,t} = (\kappa_{j,t}, Y_{j,t}, L_{j,t}, R_{j,t})$, the corresponding 'Fréchet transformed' and censored one is denoted by $C_{j,t}^{\chi} = (\tilde{\kappa}_{j,t}, \tilde{X}_{j,t}, \tilde{L}_{j,t}, \tilde{R}_{j,t})$, and, letting by convention $\mathcal{T}_j^{\chi}(\mathtt{NA}) = \mathtt{NA}$, it is defined as follows (see Figure 2)

- $$(\tilde{\kappa}_{j,t}, \tilde{X}_{j,t}) = \begin{cases} (3, \mathtt{NA}) & \text{if } \kappa_{j,t} = 1 \text{ and } Y_{j,t} < v_j\,, \\ (0, \mathtt{NA}) & \text{if } \kappa_{j,t} = 2 \text{ and } L_{j,t} < v_j\,, \\ (\kappa_{j,t}, X_{j,t}) & \text{otherwise.} \end{cases}$$

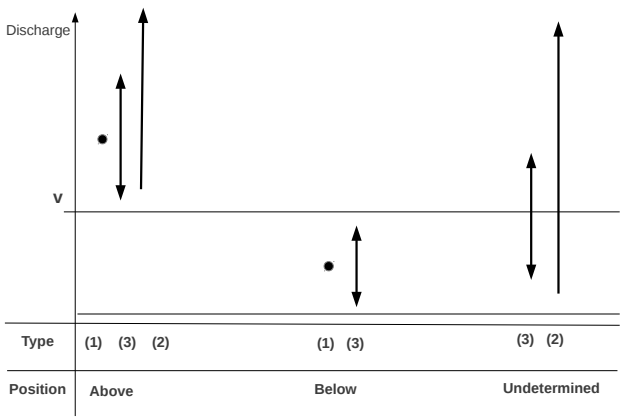

FIGURE 1. Position of marginal data points with respect to a marginal threshold v (horizontal line). Black dots: marginal data points of type $\kappa = 1$; vertical arrows: data of type $\kappa \in \{2, 3\}$.

- $\tilde{L}_{j,t} = \begin{cases} 0 & \text{if } L_{j,t} < v_j\,, \\ \mathcal{T}_j^{\chi}(L_{j,t}) & \text{otherwise.} \end{cases}$
- $\tilde{R}_{j,t} = \begin{cases} u_j & \text{if } R_{j,t} < v_j\,, \\ \mathcal{T}_j^{\chi}(R_{j,t}) & \text{otherwise.} \end{cases}$

Note that a marginal parameter χ is admissible (in view of the observations) only if all the $Y_{j,t}$'s, $R_{j,t}$'s and $L_{j,t}$'s appearing in $\mathbf{O}_t$ as argument of $\mathcal{T}_j^{\chi}$, belong to the the Pareto domain defined by χ.

By this construction, observations with undetermined position with respect to $\mathbf{v}$ have all their marginal lower bounds $\tilde{L}_{j,t}$ set to 0. They correspond to events of type $N([\frac{t}{n}, \frac{t+1}{n}) \times [\mathbf{0}, \frac{\tilde{\mathbf{R}}_t}{n}]^c) = 0$, where N is the empirical point process. The latter formulation allows to take them into account in the Poisson likelihood (section 3.3), instead of recording them as missing, which would induce a bias in favour of the largest records. In practice, if the censoring process involves a small number of censoring bounds, such situations generally occur more than once for a given upper bound $\mathbf{R}_t$.

3.3. **Poisson likelihood with censored and missing data.** In presence of censored or missing components, the likelihood of the Poisson process involves partial integration of the exponent measure λ along the axes of the Cartesian coordinate system, the integration being performed in the direction of the missing or censored coordinates. The case where some coordinates are missing (not observed at all) is easily covered: let $\mathscr{D} = \{j_1, \ldots, j_r\}$ be the non missing coordinates ($r < d$). Then, the exponent measure has a density with respect to the Lebesgue measure on the vector space spanned by $\mathscr{D}$, which is obtained by integration of (2.11) in the missing directions $\mathscr{D}_0 = \{s_1, \ldots, s_{d-r}\} = \{1, \ldots, d\} \setminus \mathscr{D}$ with integration bounds set to $(0, \infty)$. Here, the integral has an analytic expression, which is, from (2.8) and (2.9),

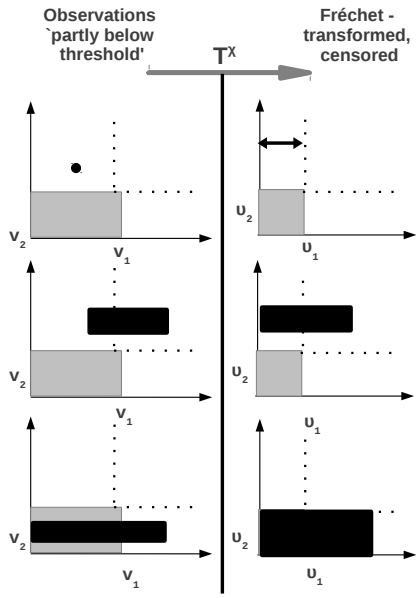

FIGURE 2. Example in the two-dimensional case of marginal transformation and censoring below threshold, for data of marginal types $(1,1)$, (upper panel) $(3,1)$ (middle panel) and $(3,3)$ (lower panel). In these three cases, observations are represented in black, respectively by a dot, an horizontal arrow and a rectangle. The Grey areas represent respectively the multivariate threshold $\mathbf{v}$ (left side) and the Fréchet transformed one $\boldsymbol{u}$ (right side). The two upper panels correspond to observations above threshold, while the lower panel shows an undetermined observation.

$$\begin{aligned}\frac{\partial \lambda_\psi(\mathbf{x})}{\partial x_{j_1}\cdots\partial x_{j_r}} &= \int_{\{\mathbf{z}:z_j=x_j(j\in\mathscr{D}),z_s\in\mathbb{R}^+(s\in\mathscr{D}_0)\}} \frac{\mathrm{d}\lambda_\psi}{\mathrm{d}\mathbf{x}}(\mathbf{z})\,\mathrm{d}z_{s_1},\ldots,\,\mathrm{d}z_{s_{d-r}} \\ &= r\sum_{m=1}^{k}\left(\frac{p_m^0\Gamma(\nu_m^0)}{\prod_{j\in\mathscr{D}}\Gamma(\nu_m^0\mu_{j,m}^0)}\prod_{j\in\mathscr{D}} x_j^{\nu_m^0\mu_{j,m}^0-1}\left(\sum_{j\in\mathscr{D}}x_j\right)^{-(\nu_m^0+1)}\right),\end{aligned} \tag{3.1}$$

with

$$\begin{gathered}\nu_m^0=\nu_m(1-\sum_{s\in\mathscr{D}_0}\mu_{s,m})\,,\quad \boldsymbol{\mu}^0_{\cdot,m}=(1-\sum_{s\in\mathscr{D}_0}\mu_{s,m})^{-1}\boldsymbol{\mu}_{\cdot,m}\,,\\ p_m^0=\frac{d}{r}(1-\sum_{s\in\mathscr{D}_0}\mu_{s,m})p_m\,.\end{gathered} \tag{3.2}$$

This is the spectral measure associated with another angular Dirichlet mixture distribution on $\mathbf{S}_r$ with parameter $\psi^0=(\nu_{1:k}^0,\boldsymbol{\mu}_{1:k}^0,p_{1:k}^0)$.

For the general case, let $(t_i\,,\ i \in \{1,\dots,n_{\mathbf{v}}\})$ denote the temporal index set of observations above threshold (at least for one component) and let $\mathbf{C}^{\chi} = \{\mathbf{C}^{\chi}_{t_i}\}(i \in \{1,\dots,n_{\mathbf{v}}\})$ be the corresponding 'Fréchet transformed' and censored data as above. The number of undetermined data is $n'_{\mathbf{v}} = \sum_{i=1}^{\mathcal{I}'} n'_i$ where $\mathcal{I}'$ is the number of undetermined temporal blocks, and n'_i is the number of undetermined observations in the ith block, *i.e.* with common upper bound $\mathbf{R}_{t'_i}$. In the sequel, for $m > 0$, the set $\mathbf{E} \setminus [\mathbf{0}, \frac{1}{m}\tilde{\mathbf{R}}_{t'_i}] = A_{\tilde{\mathbf{R}}_{t'_i},m}$ is denoted $A'_{i,m}$ and, by analogy with $A_{\boldsymbol{u}}$, the set $E \setminus [\mathbf{0}, \tilde{\mathbf{R}}_{t'_i}]$ is simply called A'_i. Since the Poisson process is exchangeable, there is no loss of generality in assuming that such blocks are formed of consecutive dates $(t'_i, \dots, t'_i + n'_i - 1)$. Finally, let $n''_{\mathbf{v}}$ be the number of observations below threshold. The number of non missing days is thus

$$n_{\text{obs}} = n_{\mathbf{v}} + n'_{\mathbf{v}} + n''_{\mathbf{v}}\,,$$

and the number of data with determined position with respect to $\mathbf{v}$ is

$$n_{\text{det}} = n_{\mathbf{v}} + n''_{\mathbf{v}}\,.$$

Let $\{j_1(t_i)\dots, j_{r(i)}(t_i)\}$ be the non-missing components $(\tilde{\kappa}_{j,t} \neq 0)$ in $\mathbf{O}_{t_i}$. The censored likelihood is

$$\begin{aligned}\mathcal{L}_{\mathbf{v}}(\mathbf{O},\theta) = \exp\Big[-n_{\text{det}}\,\lambda_\psi(A_{\boldsymbol{u}}) - \sum_{i=1}^{\mathcal{I}'} n'_i \lambda_\psi(A'_i)\Big] \times \cdots \\ \cdots \prod_{i=1}^{n_{\mathbf{v}}} \Big\{ \int_{\left[\mathscr{C}^{\chi}_{t_i}\right]^{-1}(\mathbf{C}^{\chi}_{t_i})} \frac{\partial^{r(i)}\lambda_\psi}{\partial x_{j_1(t_i)}\cdots\partial x_{j_{r(i)}(t_i)}}(\mathbf{x})\,\mathrm{d}\ell_i \prod_{j:Y_{j,t_i}>v_j} J^{\chi}_j(y_{j,t_i})\Big\}\,,\end{aligned} \tag{3.3}$$

where $(\mathscr{C}^{\chi}_t)_t$ is the censoring process on the Fréchet scale, which transforms Fréchet points $\mathbf{x}_t$ in $A_{\boldsymbol{u}}$ into censored observations $\mathbf{C}^{\chi}_t$. Integration is performed in the censored, non-missing directions, and ℓ_i is the Lebesgue measure on the corresponding subspace of $\mathbb{R}^d$.

The exponential terms in (3.3) have been obtained as a modification of those in (2.12). Namely, n is replaced with n_{obs} and the exponential term for the 'determined' data follows from

$$\exp\Big[-\frac{n_{\text{det}}}{n_{\text{obs}}}\lambda_\psi\big(A_{\boldsymbol{u},n_{\text{obs}}}\big)\Big] = \exp\Big[-n_{\text{det}}\,\lambda_\psi\big(A_{\boldsymbol{u}}\big)\Big]\,.$$

The additional term for 'undetermined' data is obtained as

$$\begin{aligned}\mathbf{P}_\theta\left\{N\left(\left[\frac{t'_i}{n_{\text{obs}}}, \frac{t'_i + n'_i - 1}{n_{\text{obs}}}\right] \times A'_{i,n_{\text{obs}}}\right) = 0\right\} &= \exp\left(-\frac{n'_i}{n_{\text{obs}}}\lambda_\psi\Big(A'_{i,n_{\text{obs}}}\Big)\right) \\ &= \exp\big(-n'_i\,\lambda_\psi\big(A'_i\big)\big)\ .\end{aligned}$$

The dimension of integration has been reduced with expression (3.1) for partial integration in the directions of missing components $(\tilde{\kappa}_{j,t_i} = 0)$. However, no closed form is available for the remaining integration in the censored directions, nor for the exponent measures of $A_{\boldsymbol{u}}$ and the A'_i's.

3.4. **Data augmentation.** In a Bayesian context, one major objective is to generate parameter samples approximately distributed according to the posterior. In classical MCMC algorithms, the value of the likelihood is needed to define the transition kernel of the Markov chain. Evaluating the integrated likelihood $\mathcal{L}_{\mathbf{v}}(\mathbf{O},\theta)$ (either by a Gaussian quadrature method or a Monte-Carlo integration) at each step of an MCMC algorithm would dramatically slow down the execution: the dimension of integration can grow up to d, for each observation, and the shape of the integrand varies from one iteration to another. In particular, large or low values of the shape parameters $\nu_{j,m}$ in ψ induce concentration of the integrand around the centres $\mu_{\cdot,m}$ or unboundedness at the simplex boundaries. Another technical issue arises from the terms $e^{-n_{\mathrm{det}}\,\lambda_\psi(A_{\boldsymbol{u}})}$, $e^{-n'_i\,\lambda_\psi(A'_i)}$ in (2.12), which have no analytic expression.

Both problems can be addressed with data augmentation methods (see *e.g.* Tanner and Wong, 1987; Van Dyk and Meng, 2001). In what follows, $[\,\cdot\,]$ denotes the distribution of random quantities as well as their density with respect to some appropriate reference measure. Proportionality between σ-finite measures is denoted by $\propto$. Thus, $[\theta]$ is the prior density and $[\theta|\mathbf{O}] \propto [\theta]\mathcal{L}_{\mathbf{v}}(\mathbf{O},\theta)$ is the posterior. The main idea is to define an augmentation space $\mathcal{Z}$, and a probability measure $[\,\cdot\,|\mathbf{O}]_+$, on the augmented space $\Theta \times \mathcal{Z}$, conditional on the observations, which may be sampled using classical MCMC methods, and which is consistent with the 'objective' on Θ. That is, the posterior on Θ must be obtained by marginalisation of $[\,\cdot\,|\mathbf{O}]_+$,

$$[\theta|\mathbf{O}] = \int_{\mathcal{Z}} [\mathbf{z},\theta|\mathbf{O}]_+ \,\mathrm{d}\mathbf{z}\,. \tag{3.4}$$

Here, an intermediate variable $\mathbf{Z}$ is introduced, so that the full conditionals $[\mathbf{Z}|\theta,\mathbf{O}]$ and $[\theta|\mathbf{Z},\mathbf{O}]$ can be directly simulated as block proposals in a *Metropolis-within-Gibbs* algorithm (Tierney, 1994). To ensure the marginalisation condition (3.4), an appropriate function $\varphi(\mathbf{z})$ is defined so that the augmented density be of the form

$$[\mathbf{z},\theta|\mathbf{O}]_+ \propto [\mathbf{z},\mathbf{O}|\theta]\varphi(\mathbf{z})\,[\theta]\,.$$

In this context, (3.4) is equivalent to

$$\mathcal{L}_{\mathbf{v}}(\mathbf{O},\theta) \propto \int [\mathbf{z},\mathbf{O}|\theta]\varphi(\mathbf{z})\,\mathrm{d}\mathbf{z} \tag{3.5}$$

In the end, a posterior sample from $[\theta|\mathbf{O}]$ is simply obtained by ignoring the $\mathbf{Z}$-components from the one produced with the 'augmented' Markov chain.

Here, the augmented data is $\mathbf{Z} = (\{Z_{j,t_i}\}_{i\le n_{\mathbf{v}}}, \mathbf{Z}'_{\boldsymbol{u}}, \{\mathbf{Z}'_i\}_{i\le\mathcal{I}'})$, where the Z_{j,t_i}'s replace the censored X_{j,t_i}'s and where the $\mathbf{Z}'$'s are the points of independent Poisson processes which account for the exponential terms $e^{-n_{\mathrm{det}}\,\lambda_\psi(A_{\boldsymbol{u}})}$ and $e^{-n'_i\,\lambda_\psi(A'_i)}$, $(i \le \mathcal{I}')$. The joint density will factorise as

$$[\mathbf{z},\mathbf{O}|\theta] = \prod_{i=1}^{n_{\mathbf{v}}} \left\{[\mathbf{z}_{t_i},\mathbf{O}_{t_i}|\theta]\right\} [\mathbf{z}'_{\boldsymbol{u}}|\theta] \prod_{i=1}^{\mathcal{I}'} [\mathbf{z}'_i|\theta]\,,$$

and the functional φ will be of the form

$$\varphi(\mathbf{z}) = \varphi_{\boldsymbol{u}}(\mathbf{z}'_{\boldsymbol{u}}) \prod_{i=1}^{\mathcal{I}'} \varphi'_i(\mathbf{z}'_i)\,.$$

Let us get into details. In a first step, let us forget about the missing components in the censored data above threshold, *i.e.*, assume that $\{(j,t_i):\ \kappa_{j,t_i}=0\}=\varnothing$. For $i \le n_{\mathbf{v}}$, let $\mathscr{D}_c(i) = (j'_1(i),\dots,j'_c(i))$ be the set of non-missing, censored coordinates in observation $\mathbf{C}^{\chi}_{t_i}$ and $\mathscr{D}_1(i) = \mathscr{D}(i) \setminus \mathscr{D}_c(i)$ be the 'exactly observed' coordinates ($\tilde{\kappa}_{j,t_i} = 1$). Let us introduce the latent variables $\mathbf{Z}_{t_i} = (Z_{j'_1,t_i},\dots,Z_{j'_c,t_i})$ that 'complete' the censored coordinates in $\mathbf{C}^{\chi}_{t_i}$. More formally, let

$$\bar{\mathbf{X}}_{t_i} \sim \frac{1}{\lambda_\psi(A_{\boldsymbol{u}})}\, \mathbb{1}_{A_{\boldsymbol{u}}}(\,\cdot\,)\, \lambda_\psi(\,\cdot\,)\,,$$

be an uncensored d-dimensional variable with Fréchet margins and dependence structure given by λ_ψ on $A_{\boldsymbol{u}}$. Then, $\mathbf{Z}_{t_i}$ is defined conditionally on the observation $\mathbf{O}_{t_i}$,

$$\begin{aligned}\big[\mathbf{Z}_{t_i}|\mathbf{O}_{t_i},\theta\big] &= [(\bar{X}_{j'_1,t_i},\dots,\bar{X}_{j'_c,t_i})|\mathbf{O}_{t_i},\theta] \\ &= [(\bar{X}_{j'_1,t_i},\dots,\bar{X}_{j'_c,t_i})|\bar{\mathbf{X}}_{t_i} \in [\mathscr{C}^{\chi}_{t_i}]^{-1}(\mathbf{C}^{\chi}_{t_i}),\psi]\,.\end{aligned} \tag{3.6}$$

In practice, the full conditionals $[Z_{j,t_i}|\{Z_{s,t_i}\}_{s\neq j},\mathbf{O},\theta]$ are functions of truncated Beta distributions that can easily be sampled in a Gibbs step of the algorithm (see Appendix A). The 'completed' data point $(\mathbf{Z}_{t_i},\mathbf{O}_{t_i})$ has density on $\mathrm{Vect}(\mathscr{D}_c)\times\mathrm{Vect}(\mathscr{D}_1)$

$$[\mathbf{z}_{t_i},\mathbf{O}_{t_i}|\theta] = \frac{\mathrm{d}\lambda_\psi}{\mathrm{d}x}(\bar{\mathbf{x}}_{t_i}) \prod_{j:\kappa_{j,t_i}=1} J^{\chi}_j(y_{j,t_i})\,,$$

where

$$\bar{x}_{j,t_i} = \begin{cases} \mathcal{T}^{\chi}_j(y_{j,t_i}) & \text{if } j\in\mathscr{D}_1(i)\,,\\ z_{j,t_i} & \text{if } j\in\mathscr{D}_c(i)\,.\end{cases}$$

Now, with missing components $\mathcal{D}_0(i)=\{j:\kappa_{j,t_i}=0\}$, the integration in $\mathcal{D}_0(i)$ can be performed analytically. Consequently, the uncensored $\bar{\mathbf{X}}_{t_i}$'s are defined on the quotient spaces $\mathbf{E}/\mathcal{D}_0(i)$ and their distributions are proportional to the exponent measures λ_{ψ^0} defined in (3.1), with density $\frac{\partial^{r(i)}\lambda_\psi}{\partial x_{j_1(t_i)}\cdots\partial x_{j_{r(i)}(t_i)}}(\,\cdot\,)$. At this stage, we have treated the integral term of (3.3), since

$$\prod_{i=1}^{n_{\mathbf{v}}} \int [\mathbf{z}_{t_i},\mathbf{O}_{t_i}|\theta]\,\mathrm{d}\ell_i(\mathbf{z}_{t_i}) = \prod_{i=1}^{n_{\mathbf{v}}} \Bigg\{ \int_{[\mathscr{C}^{\chi}_{t_i}]^{-1}(\mathbf{C}^{\chi}_{t_i})} \frac{\partial^{r(i)}\lambda_\psi}{\partial x_{j_1(t_i)}\cdots\partial x_{j_{r(i)}(t_i)}}(\mathbf{x})\,\mathrm{d}\ell_i \times\cdots$$
$$\cdots \prod_{j:y_{j,t}>v_j} J^{\chi}_j(y_{j,t}) \Bigg\}\,.$$

As for the exponential factors $\exp\big[-\lambda_\psi(A_{\boldsymbol{u},n_{\mathrm{det}}})\big]$ and $\exp\big[-n'_i\,\lambda_\psi(A'_i)\big]$ ($i\le\mathcal{I}'$), the augmentation variable $\mathbf{Z}'_{\boldsymbol{u}}$ and the $\mathbf{Z}'_i$'s are defined as independent Poisson processes and their Laplace transform plays a central role in the choice of φ. Let

us define a region $E_{\boldsymbol{u},n_{\rm det}} = \{\mathbf{x} \in (\mathbb{R}^+)^d : \|\mathbf{x}\|_1 > \min_j(\frac{u_j}{n_{\rm det}})\}$, so that $A_{\boldsymbol{u},n_{\rm det}} \subset E_{\boldsymbol{u},n_{\rm det}}$. Let us choose a multiplicative constant $\tau > 0$ and define a Poisson intensity measure $\lambda'(\,\cdot\,) = \tau\,\lambda_\psi(\,\cdot\,)$.

The augmentation process and the function $\varphi_{\boldsymbol{u}}$ are defined by

$$\begin{cases} \mathbf{Z}'_{\boldsymbol{u}} \sim PRM(\lambda') \text{ on } E_{\boldsymbol{u},n_{\rm det}}\,, \\ \varphi_{\boldsymbol{u}}(\mathbf{z}'_{\boldsymbol{u}}) = (1-1/\tau)^{\mathbf{z}'_{\boldsymbol{u}}(A_{\boldsymbol{u},n_{\rm det}})}\,, \end{cases} \tag{3.7}$$

where $\mathbf{z}'_{\boldsymbol{u}}(A_{\boldsymbol{u},n_{\rm det}})$ is the number of points forming $\mathbf{Z}'_{\boldsymbol{u}}$ which fall in $A_{\boldsymbol{u},n_{\rm det}}$.

The $\mathbf{Z}'_i$'s and the φ'_i's are defined similarly, replacing $n_{\rm det}$ with n'_i and u_j with $\tilde{R}_{t'_i,j}$. A full justification and simulation details are given in Appendix B, in particular it is shown that (3.7) implies the consistency condition (3.5). Let $\{\mathbf{x}'_{\boldsymbol{u},s} = (r'_{\boldsymbol{u},s}, \mathbf{w}'_{\boldsymbol{u},s})\}_{s\le N'_{\boldsymbol{u}}}$ be the points of $\mathbf{Z}'_{\boldsymbol{u}}$ in $E_{\boldsymbol{u},n_{\rm det}}$, the density of $\mathbf{Z}'_{\boldsymbol{u}}$ over $E_{\boldsymbol{u},n_{\rm det}}$ is

$$[\mathbf{z}'_{\boldsymbol{u}} \mid \psi] = \frac{1}{N'_{\boldsymbol{u}}!}\, e^{\frac{-n_{\rm det}\,\tau\,d}{\min_{j\le d} u_j}} \prod_{s=1}^{N'_{\boldsymbol{u}}} \frac{\tau\,d}{(r'_{\boldsymbol{u},s})^2} h_\psi(\mathbf{w}'_{\boldsymbol{u},s})\,. \tag{3.8}$$

As a conclusion, the augmented density to be sampled by the MCMC algorithm is

$$[\mathbf{z},\theta|\mathbf{O}]_+ = \left([\mathbf{z}'_{\boldsymbol{u}}|\psi] \prod_{i=1}^{n'_i} [\mathbf{z}'_i|\psi] \right) \cdot (1-1/\tau)^{\mathbf{z}'_{\boldsymbol{u}}(A_{\boldsymbol{u},n_{\rm det}}) + \sum_{i=1}^{\mathcal{I}'} \mathbf{z}'_i(A'_{i,n'_i})} \cdots$$
$$\cdots \prod_{i=1}^{n_{\mathbf{v}}} \left\{ \frac{\partial^{r(i)}\lambda_\psi}{\partial x_{j_1(i)} \cdots x_{j_{r(i)}(i)}} (\bar{\mathbf{x}}_{t_i}) \prod_{j:Y_{j,t}>v_j} J_j^\chi(y_{j,t}) \right\}. \tag{3.9}$$

4. MCMC algorithm

4.1. **Principle.** The implemented algorithm is an adaptation of the one proposed by Sabourin and Naveau (2013). It is a *Metropolis-within-Gibbs* algorithm with reversible jumps (Green, 1995) between sub-models with a given number of Dirichlet mixture components. The modified algorithm described here allows to handle additional variables, namely the marginal parameters and the augmented data.

For MCMC based inference, one must be able to generate random samples of ψ and perturb a current state ψ to obtain a proposal ψ^*. For such purpose, it is convenient to re-parametrise the model, so that moment constraints on $(\mathbf{p}, \boldsymbol{\mu})$ are automatically satisfied and the new parameter space becomes an unconstrained product space. This construction facilitates the implementation of a hybrid Gibbs algorithm. The main features of this re-parametrisation are recalled in Appendix C. In the sequel, the unconstrained parameter is denoted ϕ, and the joint parameter θ denotes the concatenation of the marginal parameters and the unconstrained dependence parameter: $\theta = (\chi, \phi)$.

There are 9 different types of moves: one *marginal move* affecting the whole set of marginal parameters, one *augmentation move* for updating the augmenting data $\{Z_{j,t_i}\}$ and one *normalising move* updating the point processes $\mathbf{Z}'_{\boldsymbol{u}}, \mathbf{Z}'_i, i \le \mathcal{I}'$ involved in the computation of the exponential terms. The 6 remaining *dependence*

moves affect the dependence parameter ϕ (whence ψ, according to the mapping $\phi \mapsto \psi$ defined by (C.1) and (C.2)), and are followed by a normalising move in order to improve the chain's mixing properties.

4.2. **Details.** In this section, the proposal are denoted with a * and the current state after ι iterations as a function of ι. The proposal kernels are $Q(\theta(\iota), \cdot)$ and their density (in the case of continuous proposals) are $q(\theta(\iota), \cdot)$. The acceptance ratios are denoted α. Additional notations distinguishing these objects defined in each type of move are omitted.

For each particular move, the acceptance ratio is set in such a way, that the unnormalised posterior measure $[\mathbf{Z}, \theta|\mathbf{O}]_+$ defined in (3.9) on the augmented parameter space be invariant under the MCMC transition kernel. In a preliminary step, likelihood optimisation is performed in the independent model (the likelihood for one multivariate observation is the product of d Pareto densities). This provides starting values for the marginal parameters as well as a Hessian matrix $\mathcal{H}$, that may be used as the inverse of a reference covariance matrix when updating the marginal parameters.

4.2.1. *Marginal moves.* The marginal parameter χ is updated as a block: The proposal is normal, with mean at $\chi(\iota)$ and co-variance matrix $\Sigma = \delta\,\mathcal{H}^{-1}$, with δ a scaling factor fixed by the user, that may typically be set around 0.5. Since the proposal density is symmetric, and since this move does not modify the dependence structure, neither the terms involving the proposal density, nor the point processes $\mathbf{Z}'$, appear in acceptance ratio. The augmented components $\bar{X}_{j,t_i} : \tilde{\kappa}_{j,t_i} \in \{2,3\}$ are left unchanged, however, those such that $\tilde{\kappa}_{j,t_i} = 1$ are updated to $\mathcal{T}^{\chi^*}_{j,t_i}(Y_{j,t_i})$. Letting $\bar{\mathbf{X}}^*_{t_i}$ denote the resulting updated data points, the acceptance ratio is

$$\alpha = \frac{[\chi^*]}{[\chi(\iota)]} \prod_{i=1}^{n_{\mathbf{v}}} \left\{ \frac{\partial^{r(i)}\lambda_\psi}{\partial x_{j_1(i)} \cdots \partial x_{j_{r(i)}(i)}}(\bar{\mathbf{X}}^*_{t_i}) \left[\frac{\partial^{r(i)}\lambda_\psi}{\partial x_{j_1(i)} \cdots \partial x_{j_{r(i)}(i)}}(\bar{\mathbf{X}}_{t_i}(\iota)) \right]^{-1} \prod_{j:Y_{j,t_i}>v_j} \frac{J_j^{\chi^*}(Y_{j,t_i})}{J_j^{\chi(\iota)}(Y_{j,t_i})} \right\}.$$

4.2.2. *Augmentation moves.* The augmented components $\{Z_{j,t_i}\} = \{\bar{\mathbf{X}}_{j,t_i} : j \in \mathscr{D}_c(i)\}$ (*c.f.* (3.3)) are directly re-sampled, one coordinate at a time, from their conditional distribution, given the other coordinates. More details about the sampling procedure are gathered in appendix A. Again, the proposal parameters are directly sampled from their full conditional distribution, and no other term in $[\cdot]_+$ is affected by the move, so that the acceptance ratio is set to $\alpha = 1$.

4.2.3. *Normalising moves.* During this move, point processes

$$\mathbf{Z}' = \left\{ \mathbf{Z}'^*_{\boldsymbol{u}}, \mathbf{Z}'_i,\ i \leq \mathcal{I}' \right\} ,$$

as defined in the end of Section 3.3, are sampled. Their conditional distribution depends on ψ only. The acceptance ratio is $\varphi(\mathbf{Z}'^*)/\varphi(\mathbf{Z}'(\iota))$, *i.e.*

$$\alpha = (1 - 1/\tau)^{\left[(\mathbf{Z}'_{\boldsymbol{u}})^*(A_{\boldsymbol{u},n_{\mathrm{det}}}) - \mathbf{Z}'_{\boldsymbol{u}}(\iota)(A_{\boldsymbol{u},n_{\mathrm{det}}})\right] + \sum_{i \leq \mathcal{I}'} \left[(\mathbf{Z}'_i)^*(A'_{i,n'_i}) - \mathbf{Z}'_i(\iota)(A'_{i,n'_i})\right]} .$$

4.2.4. *Dependence moves.* These types of moves allow to update $\phi(\iota)$ (and consequently also the original dependence parameter $\psi(\iota)$. As in Sabourin and Naveau (2013), six dependence moves are defined: μ-moves, ν-moves, e-moves, split-moves, combine-moves and shuffle moves. The only difference is that there is not enough angular data to be used during the μ-moves and the split-moves as Dirichlet kernels for the proposal distribution of some candidate $\boldsymbol{\mu}^*_{\cdot,m}$ (where $m = k(\iota)$ for a split move). Indeed, most of the observations have at least one missing or censored coordinate, so that no 'angle' is available. Consequently, the proposal is simply a Dirichlet distribution with mode at $\boldsymbol{\mu}_{\cdot,m}(\iota)$, with re-centring parameter $0 < \epsilon < 0.5$:

$$q(\boldsymbol{\mu}_{\cdot,m}(\iota), \cdot) = \mathrm{diri}_{\frac{d}{\epsilon},\gamma^*}(\cdot),$$

with $\gamma^* = (1-\epsilon)\,\boldsymbol{\mu}_{\cdot,m} + \epsilon\,(\frac{1}{d}, \ldots, \frac{1}{d})$. Each of these moves (except the shuffle move which only affects the representation of the angular distribution) is systematically followed by a normalising move, which improves the chain's mixing properties. This also avoids the computation of the 'costly' term involving the density of the points forming $\mathbf{Z}'$ in the likelihood ratio: indeed, the acceptance ratio for the two consecutive moves (dependence move with proposal kernel q, see Sabourin and Naveau (2013) for details, followed by a normalising move) is

$$\begin{aligned}\alpha = \frac{[\phi^*]}{[\phi(\iota)]}\frac{q(\phi^*, \phi(\iota))}{q(\phi(\iota), \phi^*)} \times \cdots \\ \cdots (1-1/\tau)^{\left\{\left[(\mathbf{Z}'_{\boldsymbol{u}})^*(A_{\boldsymbol{u},n_{\mathrm{det}}}) - \mathbf{Z}'_{\boldsymbol{u}}(\iota)(A_{\boldsymbol{u},n_{\mathrm{det}}})\right] + \sum_{i\leq\mathcal{I}'}\left[(\mathbf{Z}'_i)^*(A'_{i,n'_i}) - \mathbf{Z}'_i(\iota)(A'_{i,n'_i})\right]\right\}} \times \cdots \\ \cdots \prod_{i=1}^{n_{\mathbf{v}}}\left\{\frac{\partial^{r(i)}\lambda_{\psi^*}}{\partial x_{j_1(i)}\cdots\partial x_{j_{r(i)}(i)}}(\bar{\mathbf{X}}_{t_i}(\iota))\left[\frac{\partial^{r(i)}\lambda_{\psi(\iota)}}{\partial x_{j_1(i)}\cdots\partial x_{j_{r(i)}(i)}}(\bar{\mathbf{X}}_{t_i}(\iota))\right]^{-1}\right\}.\end{aligned}$$

5. Simulation example

5.1. **Experimental setting.** This section illustrates the methodology with simulated data. The purpose is not to study in full generality model features such as *e.g* the prior influence on the estimates, or the convergence on the estimates with the sample size, which are likely to be quantitatively most dependent on the censoring process. Instead, keeping in mind the application, the aim is to verify that the algorithm provides reasonable estimates with a data set which 'resembles' the particular one which motivated this work.

To allow comparison with the default space-independent framework, Bayesian inference is also performed in the independent model defined as follows: the marginal models are the same as in (2.6), the $Y_{j,t}, 1 \leq j \leq 4$ are assumed to be independent and the shape parameters $\xi_j, 1 \leq j \leq 4$ are equal to each other. A Monte-Carlo sampling scheme is straightforwardly implemented, following the same pattern as defined in the marginal moves for the dependent model. Preliminary likelihood maximisation is performed on the hydrological data, again assuming independence between locations and imposing a common shape parameter (this latter hypothesis not being rejected by a likelihood ratio test). Then, data is simulated in the model

(2.7), with marginal parameters and threshold excess probabilities (for daily data) approximately equal to the inferred ones, *i.e.*

$$\boldsymbol{\zeta} \approx (0.021, \ldots, 0.021)\,, \quad \xi = 0.4\,, \quad \log(\sigma) = (4.8, 4.6, 5.9, 5.1)\,,$$

for a total number of days $n = 148\,401$, which is the total number of days in the original data. The four-variate dependence structure is chosen as a Dirichlet mixture distribution h_ψ, where

$$\psi: \quad \mathbf{p} = (0.25, 0.25, 0.5)\,,\ \boldsymbol{\mu} = \begin{pmatrix} 0.1 & 0.7 & 0.1 \\ 0.1 & 0.1 & 0.4 \\ 0.1 & 0.1 & 0.4 \\ 0.7 & 0.1 & 0.1 \end{pmatrix}\,,\ \boldsymbol{\nu} = (70, 50, 80)\,.$$

A radial threshold for simulation on the Fréchet scale and the number of points simulated above the latter are respectively set to $r_s = -1/\log(1 - \max(\boldsymbol{\zeta}))$, $n_{\text{rad.exc}} = n * 4/r_s$. The remaining $n - n_{\text{rad.exc}}$ points are arbitrarily scattered below the radial threshold, so that the empirical probability of a radial excess is $4/r_s$, the exponent measure of the region $\{\mathbf{x} \in \mathbb{R}^4 : \|x\|_1 > r_s\}$.

Afterwards, the data set is censored following the real data's censoring pattern , *i.e.* censoring occurs on the same days and on the same locations as for the real data, with same censoring bounds (on the Fréchet scale), so that the data is observed only if the censoring threshold is exceeded. Finally, in order to account for the loss of information resulting from time dependence within the real data (whereas the simulated data are time independent), only n_{eff} data out of n are kept for inference, where $n_{\text{eff}} = \lfloor n/\text{mean cluster size} \rfloor = 118\,911$ (see Section 6 for an explanation about clusters). It should be noted that the vast majority of data points are either censored or below the multivariate threshold. Here, the threshold is arbitrarily set to the same value as the one defined for real data, *i.e.* $\mathbf{v} = (300, 320, 520, 380)$, yielding approximately equal probabilities of an excess in each direction. Only $n_{\mathbf{v}} = 162$ points are above threshold $\mathbf{v}$, from which only 6 have all their coordinates of type 1 (exact data). It should be noted that, in such a context, a simplified inferential framework in which censored data would be discarded is not an option: only 6 points would be available for inference. To fix ideas, Figure 3 shows the allocation of the number of exact marginal coordinates after censoring below threshold: most of the extracted data have only one exact coordinate.

In addition to data above threshold, the number of undetermined blocks (made of data which position with respect to the threshold is unknown) is $\mathcal{I}' = 39$, with block sizes varying between 1 and 39 845, and a total number of undetermined days $n'_{\mathbf{v}} = 112\,676$.

The bi-variate projections of the resulting data set are displayed in Figure 4.

5.2. **Inference and results.** The prior on the Dirichlet mixture distributions is specified in a similar way as in Sabourin and Naveau (2013). The number k of mixture components has truncated geometric distribution, $[k] \propto \left(1 - \frac{1}{\lambda}\right)^{k-1} \frac{1}{\lambda} \mathbb{1}_{[1,k_{\max}]}(k)$ with upper bound $k_{\max} = 10$ and mean parameter $\lambda = 4$. Also, for the sake of simplicity, all the marginal parameters are assumed to be *a priori* independent, with normal distributions (after log-transformation of the scales). The shape parameter

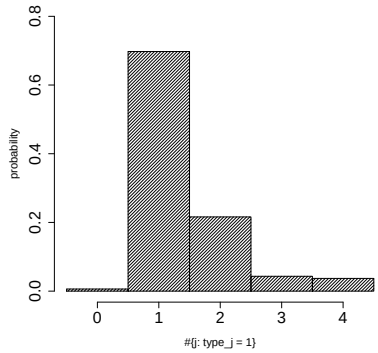

FIGURE 3. Distribution of the number of exactly observed coordinates above threshold in the simulated data set ($\#\{j : \kappa_{t_i,j} = 1\}$).

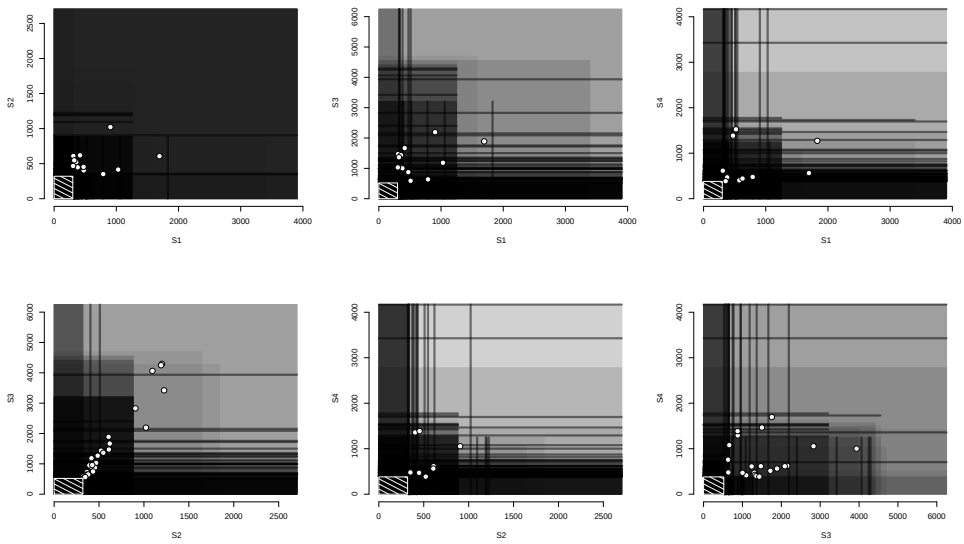

FIGURE 4. Bi-variate projection of the simulated data set after censoring. White points correspond to pairs for which both coordinates are observed. Superimposed Gray rectangles (resp. segments) represent pairs for which both coordinates (resp. one coordinate) are (is) censored. The white striped rectangle is the region below the multivariate threshold $\mathbf{v}$.

has standard normal distribution and the logarithms of the scales have mean and standard deviation both equal to 5.

As for the augmentation Poisson process data, the multiplicative constant τ involved in the Poisson intensity is set to 50. It appeared that smaller values of τ (close to 1) considerably affected the mixing properties of the chains.

Convergence of the dependence parameters $\psi(\iota)$ can be monitored using functionals based on integration of the simulated densities against Dirichlet test functions (see Sabourin and Naveau, 2013, for details). To detect possible mixing defects, six chains of 10^6 iterations each are run in parallel. Standard convergence diagnostic tests are implemented in R (Heidelberger and Welch, 1983; Gelman and Rubin, 1992), respectively testing for non-stationarity and poor mixing. The stationarity test detects three non-stationary chains out of six. The mixing properties of the three retained ones, as measured by a variance ratio inter/intra chains, are satisfying enough: all the potential scale reduction factors (Gelman and Rubin, 1992) are below 1.1.

Figure 5 displays the posterior predictive for bi-variate marginalisations of the angular density (obtained *via* (3.2)), together with the true density and posterior credible sets around the estimates. The estimated density captures reasonably well the features of the true one and the posterior quantiles are rather concentrated around the estimates. This is all the more satisfying that, at first view (Figure 4), the censored data set used for estimation seems to convey little information about the distribution of the angular components.

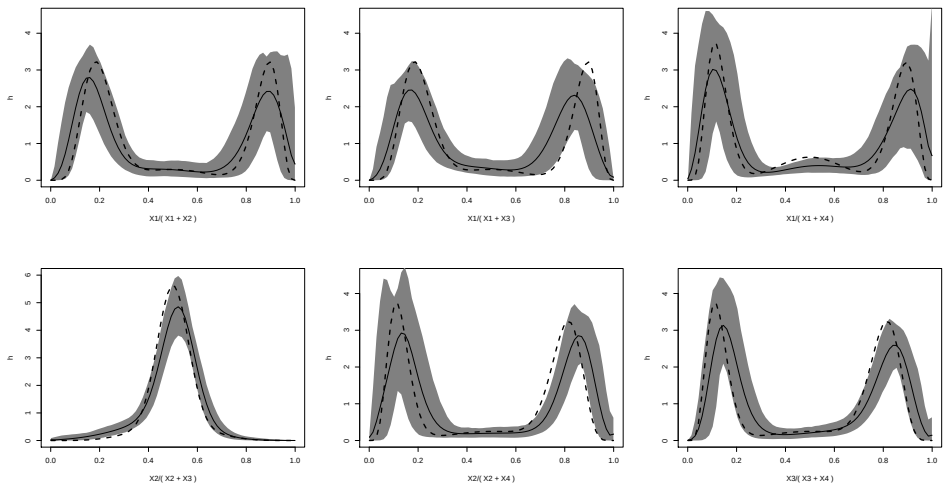

FIGURE 5. Bi-variate angular predictive densities (thick black line) with posterior credible sets (Gray regions) corresponding to the posterior $0.05 - 0.95$ quantiles, together with the true angular density (dashed line).

As for the marginal parameter components $(\chi(\iota))_\iota$, the above mentioned stationarity test detects only one non-stationary chain and the potential scale reduction factors from the Gelman tests are also satisfying. Figure 6 shows the posterior histogram of each marginal parameter, together with the prior density and the posterior histogram obtained in the independent model.

The marginal estimates obtained under the spatial independence assumption are less accurate than those obtained in the correctly specified model. Also, the observed over-estimation of the scales, combined with an under-estimation of the

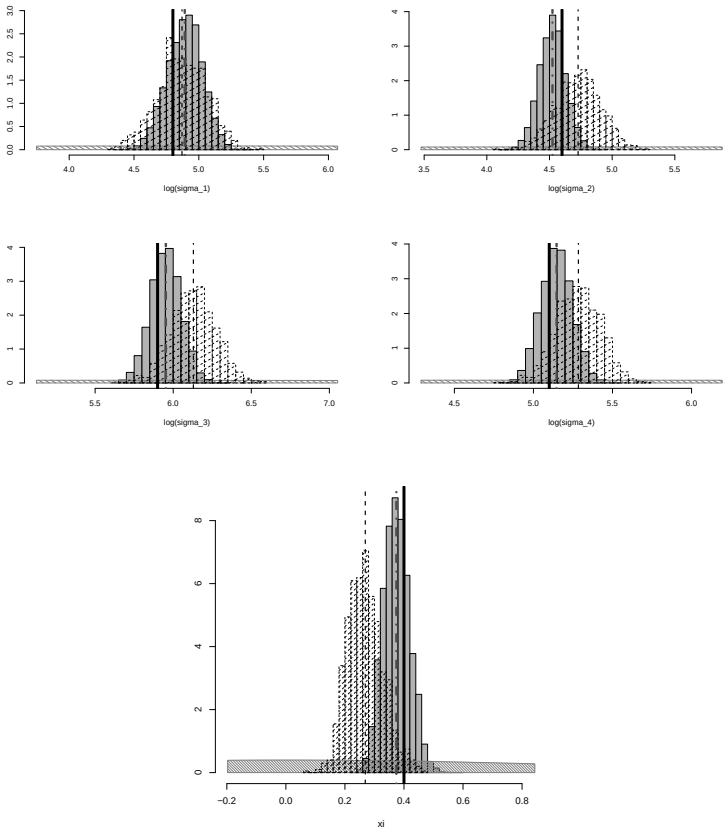

FIGURE 6. Inference of marginal parameters: logarithms of the scale parameters σ_j (four upper panels) and common shape parameter (lower panel). Gray bars and dash-dotted Gray vertical lines, posterior histograms and posterior means using the Dirichlet mixture model; black dashed bars and vertical lines: *idem* in the independent model; black thick vertical lines: true values; Gray striped area: prior density.

shape, corroborates the well known fact (Ribereau et al., 2011) that the maximum likelihood estimates of these two parameters are negatively correlated.

For some applications, including hydrological risk analysis, return level plots (*i.e.* quantile plots) are used to summarise the marginal estimates of extremes. The return level Q for a return period T at location j where data are distributed according to F_j^{χ} and where there is no temporal dependence, is usually defined as the $1 - 1/T$-quantile of F_j^{χ}. Figure 7 compares the return levels obtained in the dependent and the independent models, together with the true ones. The posterior estimates in the dependent model are very close to the truth, relatively to to the size of the credible intervals. In contrast, estimation in the independent model under-estimates the return levels, and the true curve lies outside the posterior quantiles at two locations out of four.

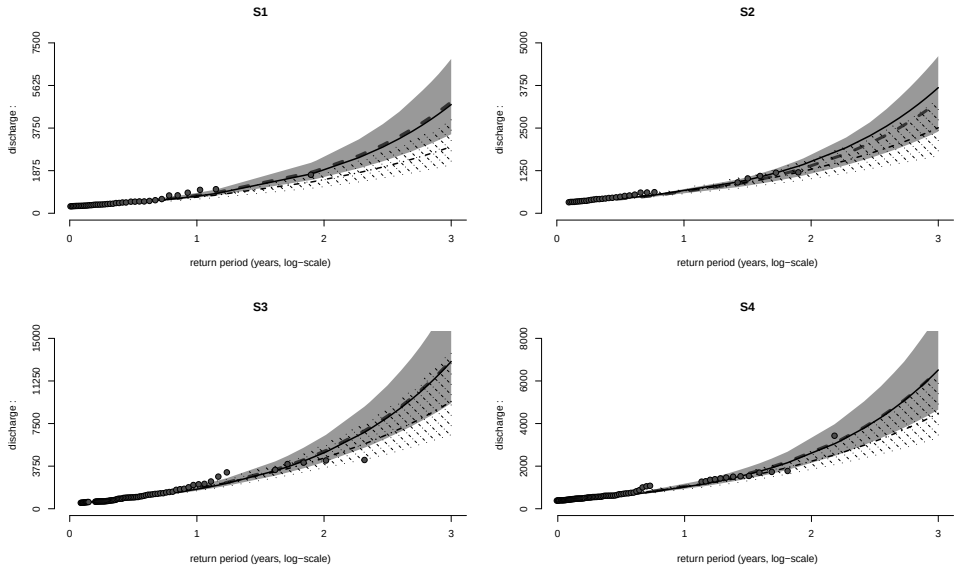

FIGURE 7. Return level plots: Quantile versus logarithm (base 10) of the return period at the four locations. Grey points: empirical return level of observed threshold excesses; Black line: true curves; Gray dashed line and shaded area: posterior mean and $0.05 - 0.95$-quantiles in the dependent model with Dirichlet mixture angular structure; Black dash-dotted line and area: *idem* in the independent model.

6. Conclusion

In this work, a flexible semi-parametric Bayesian inferential scheme is proposed to estimate the joint distribution of excesses above multivariate high thresholds, when the data are censored. A simulation example is designed on the same pattern as a real case borrowed from hydrology. Although the tuning of the MCMC algorithm requires some care, taking into account all kinds of observations for various censoring bounds allows to obtain satisfying estimates, despite the loss of information relative to the angular structure induced by the censoring process.

The main methodological novelty consists in taking advantage of the conditioning and marginalising properties of the Dirichlet distributions, in order to simulate augmentation data which 'replace' the missing ones. Also, exponential terms in the likelihood with no explicit expressions are handled by sampling well chosen functionals of augmentation Poisson processes. This new inferential framework might open the road to statistical analysis of the extremes of data sets that would otherwise have been deemed unworkable.

In the present paper, temporal dependence is not treated. In practice, time-dependent series may still be analysed using 'declustering' methods, which allow to fit the model to cluster maxima instead of the raw data. This classical approach assumes that only short-term dependence is present, *i.e.* that a condition of mixing type holds (condition D, see *e.g.* Leadbetter, 1983), so that cluster maxima may be treated as independent observations. In particular, this is the approach adopted in a forthcoming paper analysing data from the Gardons (France).

Acknowledgements

Part of this work has been supported by the EU-FP7 ACQWA Project (`www.acqwa.ch`), by the PEPER-GIS project, by the ANR-MOPERA project, by the ANR-McSim project and by the MIRACCLE-GICC project. The author would like to thank Benjamin Renard for providing the hydrological data that motivated this work and for his useful comments, and Anne-Laure Fougères and Philippe Naveau for their advice and interesting discussions we had during the writing of this paper.

Appendix A. Conditional distribution of augmented data

In this section, the full conditional distribution of the augmented variables Z_{j,t_i} ($\tilde{\kappa}_{j,t_i} \in \{2,3\}$) is derived, given all the other variables, for an excess $\mathbf{O}_{t_i}$ over threshold. Recall the Z_{j,t_i}'s are defined in (3.6) as conditional components of a random vector $\bar{\mathbf{X}}_{t_i}$ with Fréchet margins and dependence structure determined by ψ, conditionally to the observation $\mathbf{O}_{t_i}$.

Let us assume in a first step that no coordinate is missing ($\tilde{\kappa}_{j,t} \neq 0, \forall j \leq d$). The full conditional distribution is

$$[Z_{j,t_i}|\mathbf{O}_{t_i}, \{Z_{s,t_i}\}_{s\neq j}, \theta] \stackrel{d}{=} [\bar{X}_{j,t_i}[\{\bar{X}_{s,t_i}, s \neq j\}, \psi]\,.$$

In the remaining of the proof, we omit the temporal index t_i. Suppose that ψ is a mixture of k Dirichlet distributions, as in (2.9). For any bounded, continuous function g defined on $[\tilde{L}_j, \tilde{R}_j]$, the conditional expectation of $g(Z_j)$ is, up to a multiplicative constant,

$$
\begin{aligned}
\mathbb{E}\left[g(Z_j)\,|\mathbf{O}, \{Z_s\}_{s\neq j}, \theta\right] &= \mathbb{E}\left[g(\bar{X}_j)\,\middle|\bar{X}_j \in [\tilde{L}_j, \tilde{R}_j], \bar{X}_s = x_s\,(s \neq j),\, \psi\right] \\
&= \int_{\tilde{L}_j}^{\tilde{R}_j} g(x_j) h_\psi\Big(\frac{\mathbf{x}}{\sum_s x_s}\Big)\Big(\sum_s x_s\Big)^{-(d+1)} \mathrm{d}x_j \\
&= \sum_{m=1}^{k} p_m \underbrace{\int_{\tilde{L}_j}^{\tilde{R}_j} g(x_j) h_{\psi,m}\Big(\frac{\mathbf{x}}{\sum_s x_s}\Big)\Big(\sum_s x_s\Big)^{-(d+1)} \mathrm{d}x_j}_{I_m}\,,
\end{aligned}
\tag{A.1}
$$

where the $\tilde{R}_j, \tilde{L}_j$'s are the Fréchet-transformed, censored bounds defined in section 3.2.

Each term I_m ($m \leq k$ for a mixture of k components) is

$$
\begin{aligned}
I_m &= \gamma_m \int_{\tilde{L}_j}^{\tilde{R}_j} g(x_j) \prod_{s\leq d} x_s^{\nu_m\,\mu_{s,m}-1} \left(\sum_{s\leq d} x_s\right)^{-(d+1)-(\sum_{s\leq d}(\nu_m\,\mu_{s,m}-1))} \mathrm{d}x_j \\
&= \gamma_m \int_{\tilde{L}_j}^{\tilde{R}_j} g(x_j) \prod_{s\leq d} x_s^{\nu_m\,\mu_{s,m}-1} \left(\sum_{s\leq d} x_s\right)^{-(\nu_m+1)} \mathrm{d}x_j \\
&= \gamma_m \rho_j \int_{\tilde{L}_j}^{\tilde{R}_j} g(x_j) x_j^{\nu_m \mu_{j,m}-1} (s_j + x_j)^{-\nu_m-1}\, \mathrm{d}x_j\,,
\end{aligned}
$$

where $\gamma_m = \frac{\Gamma(\nu_m)}{\prod_{s=1}^{d}\Gamma(\nu_m\mu_{s,m})}$, $s_j = \sum_{s\neq j} x_s$ and $\rho_j = \prod_{s\neq j} x_s^{\nu_m\mu_{s,m}-1}$. Changing variable *via* $u = x_j/(x_j + s_j)$, the integration bounds are

$$
R'_j = \frac{\tilde{R}_j}{s_j + \tilde{R}_j}\,,\; L'_j = \frac{\tilde{L}_j}{s_j + \tilde{L}_j},
$$

and we have

$$
\begin{aligned}
I_m &= \gamma_m\,\rho_j \int_{L'_j}^{R'_j} g\left(\frac{s_j u}{1-u}\right)\left(\frac{s_j u}{1-u}\right)^{\nu_m\mu_{j,m}-1} (s_j + \frac{s_j u}{1-u})^{-\nu_m-1} \frac{s_j}{(1-u)^2}\,\mathrm{d}u \\
&= \gamma_m\,\rho_j\, s_j^{-\nu_m(1-\mu_{j,m})-1} \int_{L'_j}^{R'_j} g\left(\frac{s_j u}{1-u}\right) u^{\nu_m\mu_{j,m}-1}(1-u)^{\nu_m(1-\mu_{j,m})}\,\mathrm{d}u
\end{aligned}
$$

One recognises in the integrand the unnormalised density of a Beta random variable $U_m \sim \mathrm{beta}(a_m, b_m)$, with

$$
a_m = \nu_m\mu_{j,m}\,, \quad b_m = \nu_m(1-\mu_{j,m}) + 1\,;
$$

Let $\mathrm{IB}_{a,b}(x, y)$ denote the incomplete Beta function (*i.e.* the integral of the Beta density) between truncation bounds x and y. The missing normalising constant in the integrand is

$$D_m = \frac{\Gamma(a_m+b_m)}{\Gamma(a_m)\Gamma(b_m)\mathrm{IB}_{a_m,b_m}(L'_j,R'_j)}$$
$$= \frac{\Gamma(\nu_m)}{\Gamma(\nu_m\mu_{j,m})\Gamma(\nu_m(1-\mu_{j,m}))}\frac{1}{\mathrm{IB}_{a_m,b_m}(L'_j,R'_j)(1-\mu_{j,m})}$$

Finally, we have

$$I_m = C_m.D_m \int_{L'_j}^{R'_j} g(\frac{s_j u}{1-u})\, u^{\nu_m\mu_{j,m}-1}(1-u)^{\nu_m(1-\mu_{j,m})}\,\mathrm{d}u$$

with

$$C_m = (1-\mu_{j,m})\frac{\Gamma(\nu_m(1-\mu_{j,m}))}{\prod_{s\neq j}\Gamma(\nu_m\mu_{s,m})}\mathrm{IB}_{a_m,b_m}(L'_j,R'_j)\,\rho_j\, s_j^{-\nu_m(1-\mu_{j,m})-1}, \tag{A.2}$$

so that that the conditional expectation (A.1) is that of a mixture

$$\mathbb{E}\left[g(Z_j)|\mathbf{O},\{Z_{s,s\neq j}\}\right] = \sum_{m=1}^{j} p'_m\,\mathbb{E}\left[g(\frac{s_jU_m}{1-U_m})\right]$$

with weights $(p'_m)_{m\leq k}$,

$$p'_m = p_m\, C_m\,, \tag{A.3}$$

where C_m is given by (A.2).

As a conclusion, the conditional variable $[Z_j|\mathbf{O},Z_{s\neq j},\theta]$ is a mixture distribution of k components

$$\left(p'_m, V_{j,m} = \frac{s_jU_m}{1-U_m}\right)_{1\leq m\leq k}. \tag{A.4}$$

In presence of missing coordinates $\mathcal{D}_0(i)$, (A.4) still holds, up to replacing the mixture parameters $(\mathbf{p},\boldsymbol{\mu},\boldsymbol{\nu})$ with $(\mathbf{p}^0,\boldsymbol{\mu}^0,\boldsymbol{\nu}^0)$ as in (3.2), with multiplicative corrective factors $1-\sum_{j\in\mathcal{D}_0(i)}\mu_{j,m}$.

Appendix B. Consistency of the augmentation Poisson processes $\mathbf{Z}'_{\boldsymbol{u}}, \mathbf{Z}'_i$.

For (3.5) to hold, it is sufficient to define $\mathbf{Z}'_{\boldsymbol{u}}, \mathbf{Z}'_i$, $(i\leq\mathcal{I}')$, independent, with respective densities $[\mathbf{z}'_{\boldsymbol{u}}|\psi]$, $[\mathbf{z}'_i|\psi]$, together with $\varphi_{\boldsymbol{u}}(\mathbf{z}'_{\boldsymbol{u}}), \varphi_i(\mathbf{z}'_i)$, such that

$$\begin{cases}\mathbb{E}\left[\varphi_{\boldsymbol{u}}(\mathbf{Z}'_{\boldsymbol{u}})\right] = \exp\left(-n_{\mathrm{det}}\,\lambda_\psi\left(A_{\boldsymbol{u}}\right)\right)\,, \\ \mathbb{E}\left[\varphi_i(\mathbf{Z}'_i)\right] = \exp\left(-n'_i\,\lambda_\psi(A'_i)\right) \qquad (i\leq\mathcal{I}')\,.\end{cases} \tag{B.1}$$

At this point, one may be tempted to define $\mathbf{Z}'_{\boldsymbol{u}}$ as a Poisson process with intensity λ_ψ on some $E\supset A_{\boldsymbol{u},n_{\mathrm{det}}}$, and $\varphi(\mathbf{Z}_{\boldsymbol{u}})$ as the indicator $\mathbb{1}_{\mathbf{z}'_{\boldsymbol{u}}(A_{\boldsymbol{u},n_{\mathrm{det}}})=0}$, with a similar definition for the φ'_i's and the $\mathbf{Z}'_i$'s. Indeed, one would have $\mathbb{E}\left[\mathbf{1}_{\mathbf{z}'(A_{\boldsymbol{u},n_{\mathrm{det}}})=0}\right] = \exp\left(-n_{\mathrm{det}}\,\lambda_\psi\left(A_{\boldsymbol{u}}\right)\right)$, as required. However, even if this construction is valid in theory, it leads to a very large rate of rejection in the Metropolis algorithm: φ has too much variability around its mean value and the proposal is systematically rejected each time a point in the process falls in the failure region.

Here is detailed the alternative construction of $\mathbf{Z}'_{\boldsymbol{u}}$ and $\varphi_{\boldsymbol{u}}$ which is used in this paper. To wit, $\varphi_{\boldsymbol{u}}$ is a smoothed version of the above characteristic function.

Recall that, if f is a bounded, continuous function defined on a nice space E and if N is PRM(λ) on E, then the Laplace transform $Lap_N(f) \hat{=} \mathbb{E}(e^{-N(f)})$ is (Resnick, 1987, Chap. 3)

$$Lap_N(f) = \exp\left(-\int_E (1 - e^{-f(s)})\, \mathrm{d}\lambda(s)\right),$$

where, for a random measure $N = \sum_{i=1}^{N(E)} \mathbb{1}_{s_i}(\,\cdot\,)$,

$$N(f) = \int_E f\, \mathrm{d}N = \sum_{i=1}^{N(E)} f(s_i).$$

Let $E_{\boldsymbol{u},n_{\mathrm{det}}}$ be any region of $\mathbf{E}$ containing $A_{\boldsymbol{u},n_{\mathrm{det}}}$ as *e.g.* the one defined in Section 3.4. It is enough to find a function $f_{\boldsymbol{u}}$ on $E_{\boldsymbol{u},n_{\mathrm{det}}}$ as above and an intensity measure λ' such that

$$\int_{E_{\boldsymbol{u},n_{\mathrm{det}}}} (1 - e^{-f_{\boldsymbol{u}}(s)})\, \mathrm{d}\lambda'(s) = \lambda_\psi(A_{\boldsymbol{u},n_{\mathrm{det}}})\,. \tag{B.2}$$

Indeed, letting

$$\begin{cases} \mathbf{Z}'_{\boldsymbol{u}} \sim PRM(\lambda') \text{ on } E_{\boldsymbol{u},n_{\mathrm{det}}}\,, \\ \varphi_{\boldsymbol{u}}(\mathbf{z}') = \exp(-\mathbf{z}'(f_{\boldsymbol{u}}))\,, \end{cases}$$

one has $\mathbb{E}\big[\varphi(\mathbf{Z}'_{\boldsymbol{u}})|\psi\big] = Lap_{\mathbf{Z}'_{\boldsymbol{u}}}(f_{\boldsymbol{u}}) = \exp(-\lambda_\psi(A_{\boldsymbol{u},n_{\mathrm{det}}}))$, as required by (B.1).

In order to satisfy (B.2), let us fix a multiplicative constant $\tau > 1$ and define

$$\begin{cases} \lambda'(\,\cdot\,) = \tau\, \lambda_\psi(\,\cdot\,)\,, \\ f_{\boldsymbol{u}}(\mathbf{x}) = -\log(1 - 1/\tau)\mathbf{1}_{A_{\boldsymbol{u},n_{\mathrm{det}}}}(\mathbf{x})\,. \end{cases}$$

Then, (B.2) holds. The points of $\mathbf{Z}'_{\boldsymbol{u}}$ can easily be simulated (see Resnick, 1987, Chap.3): the number of points $N'_{\boldsymbol{u}}$ in $E_{\boldsymbol{u},n_{\mathrm{det}}}$ is a Poisson random variable with mean equal to

$$\lambda'(E_{\boldsymbol{u},n_{\mathrm{det}}}) \;=\; \frac{\tau\; d}{\min_j(u_j\;/n_{\mathrm{det}})}\,,$$

and each point has density in polar coordinates equal to $\frac{1}{\lambda'(E_{\boldsymbol{u},n_{\mathrm{det}}})}\frac{\tau d}{r^2} h_\psi(\mathbf{w})$.

To sum up, the desired functional $\varphi_{\boldsymbol{u}}$ is

$$\varphi_{\boldsymbol{u}}(\mathbf{z}'_{\boldsymbol{u}}) = e^{\log(1-1/\tau)\mathbf{z}'_{\boldsymbol{u}}(A_{\boldsymbol{u},n_{\mathrm{det}}})} = (1 - 1/\tau)^{\mathbf{z}'_{\boldsymbol{u}}(A_{\boldsymbol{u},n_{\mathrm{det}}})}\,,$$

as defined in (3.7). As claimed above, for τ close to one, $\varphi'_{\boldsymbol{u}}(\mathbf{z}'_{\boldsymbol{u}})$ is close to $\mathbf{1}_{\mathbf{z}'(A_{\boldsymbol{u},n_{\mathrm{det}}})=0}$.

Appendix C. Re-parametrisation of the Dirichlet mixture model for Bayesian inference

For the sake of completeness, the main features of the re-parametrisation introduced by Sabourin and Naveau (2013), are recalled below.

A Dirichlet mixture distribution with parameter ψ with k mixture components, subject to constraint (2.10) can be re-parametrised with an unconstrained parameter

$$\phi = (\boldsymbol{\mu}_{\cdot,1:k-1}, e_{1:k-1}, \nu_{1:k}) \in (\mathbf{S}_d)^{k-1} \times (0,1)^{k-1} \times (\mathbb{R}^{*+})^k,$$

where the *eccentricity parameter* $\mathbf{e} = e_{1:k-1}$ rules the relative positions of partial centres of mass $\boldsymbol{\gamma}_m$ defined by

$$\boldsymbol{\gamma}_m = \frac{1}{1 - \sum_{j=1}^{m} p_j} \sum_{j=m+1}^{k} p_j \, \boldsymbol{\mu}_{\cdot,j} \,,$$

from which the weights and the last Dirichlet kernel centre $\boldsymbol{\mu}_{\cdot,k}$ can be sequentially reconstructed. Namely, the moment constraint (2.10) says that $\boldsymbol{\gamma}_0 = \left(\frac{1}{d}, \ldots, \frac{1}{d}\right)$. Now, for $m \geq 1$, by centroid decomposition, the simplex points $\boldsymbol{\mu}_{\cdot,m}, \boldsymbol{\gamma}_{m-1}$ and $\boldsymbol{\gamma}_m$ are aligned (see Figure 8 which illustrates the bi-dimensional case), so that there exists $t \geq 0$, such that $\boldsymbol{\gamma}_m = \boldsymbol{\gamma}_{m-1} + t\,(\boldsymbol{\gamma}_{m-1} - \boldsymbol{\mu}_{\cdot,m})$.

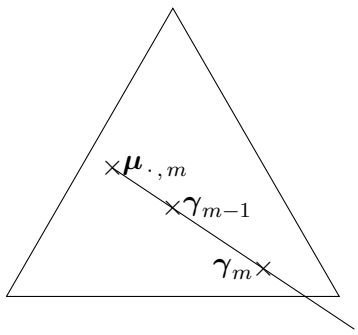

FIGURE 8. Geometric construction of the partial centres of mass on the simplex $\mathbf{S}_3$ at step m. The simplex points γ_m, γ_{m-1} and m^{th} kernel centre $\boldsymbol{\mu}_{\cdot,m}$, belong to a common line and γ_{m-1} lies between γ_m and $\boldsymbol{\mu}_{\cdot,m}$.

Let

$$L_m = \sup\left\{s \geq 0 : \; \boldsymbol{\gamma}_{m-1} + s\,(\boldsymbol{\gamma}_{m-1} - \boldsymbol{\mu}_{\cdot,m}) \in \mathbf{S}_d\right\} \,,$$

which has analytic expression $L_m = \min_{i \in \mathcal{C}_m} \frac{\gamma_{i,m-1}}{\mu_{i,m} - \gamma_{i,m-1}}$, where $\mathcal{C}_m$ is the index set $\left\{i \in \{1, \ldots, d\} \; : \; \gamma_{i,m-1} - \mu_{i,m} < 0\right\}$.

Finally, e_m determines $\boldsymbol{\gamma}_m$ *via*

$$\boldsymbol{\gamma}_m = \boldsymbol{\gamma}_{m-1} + e_m \, L_m \, (\boldsymbol{\gamma}_{m-1} - \boldsymbol{\mu}_{\cdot,m}). \tag{C.1}$$

Thus, given $\mathbf{e}$ and $\boldsymbol{\mu}_{\cdot,1:k-1}$, the positions of the $\boldsymbol{\gamma}_m$ $(1 \leq m \leq k-1)$ are known, the last centre of mass is $\boldsymbol{\mu}_{\cdot,k} = \boldsymbol{\gamma}_{k-1}$ by definition, and the weights are sequentially reconstructed *via*

$$p_m = (1 - \sum_{j<m} p_j) \frac{e_m L_m}{1 + e_m \, L_m} \,. \tag{C.2}$$

The main advantage of this geometric construction is that the parameter space $\Phi = (\mathbf{S}_d)^{k-1} \times (0,1)^{k-1} \times (\mathbb{R}^{*+})^k$ is an unconstrained product space, on which it is easy to define a prior, and which can be spanned by a MCMC algorithm with relatively simple proposal distributions. Indeed, the proposals will automatically belong to the parameter space, because there remains no constraint of type (2.10).

References

Boldi, M.-O. and Davison, A. C. (2007). A mixture model for multivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):217–229.

Coles, S. and Tawn, J. (1991). Modeling extreme multivariate events. *JR Statist. Soc. B*, 53:377–392.

Davison, A. and Smith, R. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 393–442.

Einmahl, J., de Haan, L., and Piterbarg, V. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *The Annals of Statistics*, 29(5):1401–1423.

Einmahl, J. and Segers, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics*, 37(5B):2953–2989.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

Gómez, G., Calle, M. L., and Oller, R. (2004). Frequentist and bayesian approaches for interval-censored data. *Statistical Papers*, 45(2):139–173.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711.

Guillotte, S., Perron, F., and Segers, J. (2011). Non-parametric bayesian inference on bivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Gumbel, E. (1960). Distributions des valeurs extrêmes en plusieurs dimensions. *Publ. Inst. Statist. Univ. Paris*, 9:171–173.

Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, pages 1109–1144.

Leadbetter, M. (1983). Extremes and local dependence in stationary sequences. *Probability Theory and Related Fields*, 65(2):291–306.

Ledford, A. and Tawn, J. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Neppel, L., Renard, B., Lang, M., Ayral, P., Coeur, D., Gaume, E., Jacob, N., Payrastre, O., Pobanz, K., and Vinet, F. (2010). Flood frequency analysis using historical data: accounting for random and systematic errors. *Hydrological Sciences Journal–Journal des Sciences Hydrologiques*, 55(2):192–208.

Resnick, S. (1987). *Extreme values, regular variation, and point processes, volume 4 of Applied Probability. A Series of the Applied Probability Trust.* Springer-Verlag, New York.

Resnick, S. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling.* Springer Series in Operations Research and Financial Engineering.

Ribereau, P., Naveau, P., and Guillou, A. (2011). A note of caution when interpreting parameters of the distribution of excesses. *Advances in Water Resources*, 34(10):1215–1221.

Sabourin, A. and Naveau, P. (2013). Bayesian dirichlet mixture model for multivariate extremes: A re-parametrization. *Computational Statistics & Data Analysis*, (0):–.

Schnedler, W. (2005). Likelihood estimation for censored random vectors. *Econometric Reviews*, 24(2):195–217.

Smith, R. (1994). Multivariate threshold methods. *Extreme Value Theory and Applications*, 1:225–248.

Smith, R., Tawn, J., and Coles, S. (1997). Markov chain models for threshold exceedances. *Biometrika*, 84(2):249–268.

Stephenson, A. (2003). Simulating multivariate extreme value distributions of logistic type. *Extremes*, 6(1):49–59.

Stephenson, A. (2009). High-dimensional parametric modelling of multivariate extreme events. *Australian & New Zealand Journal of Statistics*, 51(1):77–88.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728.

Van Dyk, D. and Meng, X. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50.

Université de Lyon, CNRS UMR 5208, Université de Lyon 1, Institut Camille Jordan, 43 blvd. du 11 novembre 1918, F-69622 Villeurbanne cedex, France ; (Laboratoire des Sciences du Climat et de l'Environnement, CNRS-CEA-UVSQ, 91191 Gif-sur-Yvette, France)

E-mail address: sabourin@math.univ-lyon1.fr

4.2 Analyse des données des Gardons

COMBINING REGIONAL ESTIMATION AND HISTORICAL FLOODS: A MULTIVARIATE SEMI-PARAMETRIC PEAKS-OVER-THRESHOLD MODEL WITH CENSORED DATA

A. SABOURIN AND B. RENARD

Multivariate extremes; censored data; semi-parametric Bayesian inference; mixture models; reversible-jump algorithm

Abstract. The estimation of extreme flood quantiles is challenging due to the relative scarcity of extreme data compared to typical target return periods. Several approaches have been developed over the years to face this challenge, including regional estimation and the use of historical flood data. This paper investigates the combination of both approaches using a multivariate peaks-over-threshold model, that allows estimating altogether the intersite dependence structure and the marginal distributions at each site. The joint distribution of extremes at several sites is constructed using a semi-parametric Dirichlet Mixture model. The existence of partially missing and censored observations (historical data) is accounted for within a data augmentation scheme. This model is applied to a case study involving four catchments in Southern France, for which historical data are available since 1604. The comparison of marginal estimates from four versions of the model (with or without regionalizing the shape parameter; using or ignoring historical floods) highlights significant differences in terms of return level estimates. Moreover, the availability of historical data on several nearby catchments allows investigating the asymptotic dependence properties of extreme floods. Catchments display a a significant amount of asymptotic dependence, calling for adapted multivariate statistical models.

1. Introduction

Statistical analysis of extremes of uni-variate hydrological time series is a relatively well chartered problem. Two main representations can be used in the context of extreme value theory (*e.g.* Madsen et al., 1997b; Coles, 2001): block maxima (typically, annual maxima) can be modeled using a Generalized Extreme Value (GEV) distribution (see *e.g.* Hosking, 1985), while flood peaks over a high threshold (POT) are commonly modeled with a Generalized Pareto (GP) distribution (see *e.g.* Hosking and Wallis, 1987; Davison and Smith, 1990; Lang et al., 1999). One major issue in at-site flood frequency analysis is related to data scarcity (Neppel et al., 2010): as an illustration, most of the recorded flood time series in France are less than 50 years long, whereas flood return periods of interest are typically well above 100 years. Moreover, an additional challenge arises if one is interested in multivariate extremes at several locations. A complete understanding of the joint behavior of extremes at different locations requires to model their dependence structure as well. While there exists a multivariate extreme value theory (*e.g.*

Date: September 19, 2013.

Coles and Tawn, 1991; De Haan and De Ronde, 1998), its practical application is much more challenging than with standard univariate approaches.

1.1. **Regional estimation.** In order to address the issue of data scarcity in at-site flood frequency analysis, hydrologists have developed methods to jointly use data from several sites: this is known as Regional Frequency Analysis (RFA) (*e.g.* Hosking and Wallis, 1997; Madsen and Rosbjerg, 1997; Madsen et al., 1997a). The basis of RFA is to assume that some parameters governing the distributions of extremes remain constant at the regional scale (see *e.g.* the 'Index Flood' approach of Dalrymple, 1960). All extreme values recorded at neighboring stations can hence be used to estimate the regional parameters, which increases the number of available data.

The joint use of data from several sites induces a technical difficulty: the spatial dependence between sites has to be modeled. A common assumption has been to simply ignore spatial dependence by assuming that the observations recorded simultaneously at different sites are independent, which is often unrealistic (see Stedinger, 1983; Hosking and Wallis, 1988; Madsen and Rosbjerg, 1997, for appraisals of this assumption). An alternative approach uses elliptical copulas to describe spatial dependence (Renard and Lang, 2007; Renard, 2011). While this approach allows moving beyond the spatial independence assumption, it is not fully satisfying. Indeed, such copula models are not compatible with multivariate extreme value theory (Resnick, 1987, 2007; Beirlant et al., 2004). This may alter uncertainty assessments about regional parameters (in particular for shape parameters) and, in turn, about extreme quantiles. In this context, using a dependence model compatible with multivariate extreme value theory is of interest.

1.2. **Historical data.** Beside regional analysis methods, an alternative way to reduce uncertainty is to take into account historical flood records to complement the systematic streamflow measurements over the recent period (see *e.g.* Stedinger and Cohn, 1986; O'Connel et al., 2002; Parent and Bernier, 2003; Reis and Stedinger, 2005; Naulet et al., 2005; Neppel et al., 2010; Payrastre et al., 2011). This results in a certain amount of censored and missing data, so that any likelihood-based inference ought to be conducted using a censored version of the likelihood function. Also, in a regional POT context, some observations may not be concomitantly extreme at each location, so that the marginal GP distribution does not apply to them. A 'censored likelihood' inferential framework for extremes has been introduced to take into account such observations (Smith, 1994; Ledford and Tawn, 1996; Smith et al., 1997). The information carried by partially censored data is likely to be all the more relevant in a multivariate, dependent context, where information at one well gauged location can be transferred to poorly measured ones.

1.3. **Multivariate modeling.** The family of admissible dependence structures between extreme events is, by nature, too large to be fully described by any parametric model (see further discussion in section 3.2). For applied purposes, it is common to restrict the dependence model to a parametric sub-class, such as, for example, the Logistic model and its asymmetric and nested extensions (Gumbel, 1960; Coles and Tawn, 1991; Stephenson, 2003, 2009). The main practical advantage is that the censored versions of the likelihood are readily available, but

parameters are subject to non-linear constraints and structural modeling choices have to be made *a priori*, *e.g.*, by allowing only bi-variate or tri-variate dependence between closest neighbors. An alternative to parametric modeling is to resort to 'semi-parametric' mixture models (some would say 'non-parametric' because it can approach any dependence structure): the distribution function characterizing the dependence structure is written as a weighted average of an arbitrarily large number of simple parametric components. This allows keeping the practical advantages of a parametric representation while providing a more flexible model.

1.4. **Objectives: Combining historical data and regional analysis.** Our aim is to combine regional analysis and historical data by modeling altogether the marginal distributions and the dependence structure of excesses above large thresholds at neighboring locations with partially censored data. Combined historical/regional approaches have been explored by a few authors (Tasker and Stedinger, 1987, 1989; Jin and Stedinger, 1989; Gaume et al., 2010). This paper builds on this previous work and extends it to a multivariate POT context, where each d-variate observation corresponds to concomitant streamflows recorded at d sites. This is to be compared with the multivariate annual maxima approach, where each d-variate observation corresponds to componentwise annual maxima that may have been recorded during distinct extreme episodes.

In this paper, a multivariate POT model is implemented in order to combine regional estimation and historical data. This model is used to investigate two scientific questions. Firstly, the relative impact of regional and historical information on marginal quantile estimates at each site is investigated. Secondly, the existence of historical data describing exceptional flood events at several nearby catchments provides an unique opportunity to investigate the nature and the strength of inter-site dependence at very high levels (which would not be possible using short series of systematic data only).

Multivariate POT modeling is implemented in a Bayesian, semi-parametric context. The dependence structure is described using a Dirichlet Mixture (DM) model. The DM model was first introduced by Boldi and Davison (2007), and its reparametrized version (Sabourin and Naveau, 2013) allows for Bayesian inference with a varying number of mixture components. A complete description of the model and of the reversible-jump Markov Chain Monte-Carlo (MCMC) algorithm used for inference with non censored data is given in Sabourin and Naveau (2013). The adaptation of the inferential framework to the case of partially censored and missing data is fully described from a statistical point of view in a forthcoming paper[1]. One practical advantage of this mixture model is that no additional structural modeling choice needs to be made, which allows to cover an arbitrary wide range of dependence structures. In this work, we aim at modeling the multivariate distribution of $d = 4$ locations. However, the methods presented here are theoretically valid in any dimension, and computationally realistic in moderate dimensions (say $d \leq 10$).

The remainder of this paper is organized as follows: the dataset under consideration is described in Section 2, and a multivariate declustering scheme is proposed

[1]Sabourin, A., 'Semi-parametric modelling of excesses above high multivariate thresholds with censored data', submitted, preprint available online at `http://www.lsce.ipsl.fr/Phocea/Pisp/index.php?nom=anne.sabourin`

to handle temporal dependence. Section 3 summarizes the main features of the multivariate POT model and describes the inferential algorithm. In Section 4, the model is fitted to the data and results are described. Section 5 discusses the main limitations of this study and proposes avenues for improvement, while Section 6 summarizes the main findings of this study.

2. Hydrological data

2.1. **Overview.** The dataset under consideration consists of discharge recorded in the area of the 'Gardons', in the south of France. Four catchments (Anduze, 540 km^2, Alès, 320 km^2, Mialet, 219 km^2, and Saint-Jean,154 km^2) are considered. They are located relatively close to each other (see Figure 1). Discharge data (in $m^3.s^{-1}$) were reconstructed by Neppel et al. (2010) from systematic measurements (recent period) and historical floods. Neppel et al. (2010) estimated separately the marginal uni-variate extreme value distributions for yearly maximum discharges, taking into account measurement and reconstruction errors arising from the conversion of water levels into discharge. The earliest record dates back to 1604, September 10th and the latest was made in 2010, December 31st.

In this work, since we are more interested in the dependence structure between simultaneous records than between yearly maxima, we model multivariate excesses over threshold, and the variable of interest becomes (up to declustering) the daily peakflow. Of course, most of the $N = 14\,841$ daily peaklflows are censored (*e.g.*, most historical data are only known to be smaller than the yearly maximum for the considered year). For the sake of simplicity, we do not take into account any possible measurement errors.

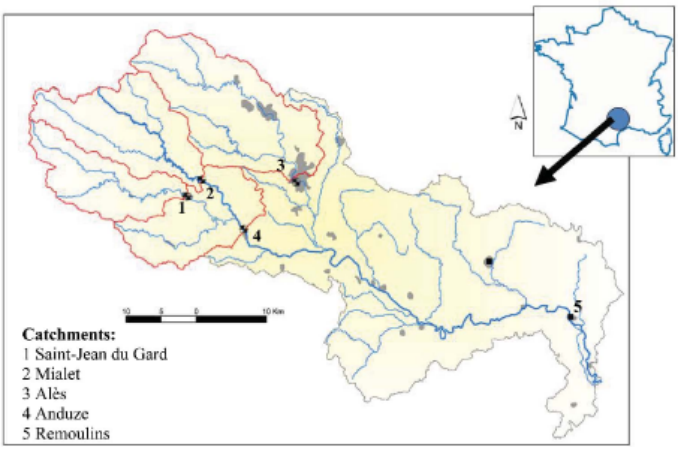

Figure 1. Hydrological map of the area of the Gardons, France (Neppel et al., 2010)

The geographic proximity of the four considered stations suggests dependence at high levels. This is visually confirmed by the pairwise plots in Figure 3, obtained after declustering (see Section 2).

The marginal data are classified into four different types, numbered from 0 to 3: '0' denotes missing data, '1' indicate an 'exact' record. Data of type '2' are right-censored: the discharge is known to be greater than a given value. Finally,

type '3' data are left- and right-censored: the discharge is known to be comprised between a lower (possibly 0) and an upper bound. Most data on the historical period are of type 3. In the sequel, j ($1 \leq j \leq d$) denotes the location index and t ($1 \leq t \leq n$) is used for the the temporal one. A marginal observation $O_{j,t}$ is a 4-uple $O_{j,t} = (\kappa_{j,t}, Y_{j,t}, L_{j,t}, R_{j,t}) \in \{0,1,2,3\} \times \mathbb{R}^3$, where κ, Y, L and R stand respectively for the data type, the recorded discharge (or some arbitrary value if $\kappa \neq 1$, which we denote `NA`), the lower bound (set to 0 if missing), and the upper bound (set to $+\infty$ if missing).

2.2. **Data pre-processing: extracting cluster maxima.** Temporal dependence is handled by declustering, *i.e.* by fitting the model to cluster maxima instead of the raw daily data. The underlying assumption is that only short term dependence is present at extreme levels, so that excesses above high thresholds occur in clusters. Cluster maxima are treated as independent data to which a model for threshold excesses may be fitted. For an introduction to declustering techniques, the reader may refer to Coles (2001) (Chap.5). For more details, see *e.g.* Leadbetter (1983), or Davison and Smith (1990) for applications when the quantities of interest are cluster maxima. Also, Ferro and Segers (2003) propose a method for identifying the optimal cluster size, after estimating the extremal index. However, this latter approach relies heavily on 'inter-arrival times', which are not easily available in our context of censored data. In this study, we adopt a simple 'run declustering' approach, following Coles and Tawn (1991) or Nadarajah (2001) : a multivariate declustering threshold $\mathbf{v} = (v_1, \ldots v_d)$ is specified (typically, $\mathbf{v} = (300, 320, 520, 380)$ respectively for Saint-Jean, Mialet, Anduze and Alès), as well as a duration τ representative of the hydrological features of the catchment (typically $\tau = 3$ days). Following common practice (Coles, 2001), the thresholds are chosen in regions of stability of the maximum likelihood estimates of the marginal parameters.

In a censored data context, a marginal data $O_{j,t}$ exceeds v_j (*resp.* is below v_j) if $\kappa_{j,t} = 1$ and $Y_{j,t} > v_j$ (*resp.* $Y_{j,t} < v_j$), or if $\kappa_{j,t} \in \{2,3\}$ and $L_{j,t} > v_j$ (*resp.* $\kappa_{j,t} = 3$ and $R_{j,t} < v_j$). If none of these conditions holds, we say that the data point has undetermined position with respect to the threshold. This is typically the case when some censoring intervals intersect the declustering thresholds whereas no coordinate is above threshold.

A cluster is initiated when at least one marginal observation $O_{j,t}$ exceeds the corresponding marginal threshold v_j. It ends only when, during at least τ successive days, all marginal observations are either below their corresponding threshold, or have undetermined position. Let $\{t_i, 1 \leq i \leq n_{\mathbf{v}}\}$ be the temporal indices of cluster starting dates. A cluster maximum $\mathbf{C}^{\vee}_{t_i}$ is the component-wise 'maximum' over a cluster duration $[t_i, \ldots, t_i + r]$. Its definition require special care in the context of censoring: the marginal cluster maximum is $C^{\vee}_{j,t_i} = \left(\kappa^{\vee}_{j,t_i}, Y^{\vee}_{j,t_i}, L^{\vee}_{j,t_i}, R^{\vee}_{j,t_i}\right)$, with $Y^{\vee}_{j,t_i} = \max_{t_i \leq t \leq t_i + r}\{Y_{j,t}\}$ and similar definitions for $L^{\vee}_{j,t_i}, R^{\vee}_{j,t_i}$. The marginal type $\kappa^{\vee}_{j,t_i}$ is that of the 'largest' record over the duration. More precisely, omitting the temporal index, if $Y^{\vee}_j > L^{\vee}_j$, then $\kappa^{\vee}_j = 1$. Otherwise, if $L^{\vee}_j < R^{\vee}_j$, then $\kappa^{\vee}_j$ is set to $= 3$; otherwise, if $L^{\vee}_j > 0$, then $\kappa^{\vee}_j = 2$; If none of the above holds, then the jth cluster coordinate is missing and $\kappa^{\vee}_j = 0$.

Figure 2 shows the uni-variate projections of the multivariate declustering scheme, at each location. Points and segments below the declustering threshold indicate situations when the threshold was not exceeded at the considered location but at another one.

Anticipating Section 3, marginal cluster maxima below threshold are censored in the statistical analysis, so that their marginal types are always set to 3, with lower bound at zero and upper bound at the threshold. This approach, fully described *e.g.* in Ledford and Tawn (1996), prevents from having to estimate the marginal distribution below threshold, which does not participate in the dependence structure of extremes.

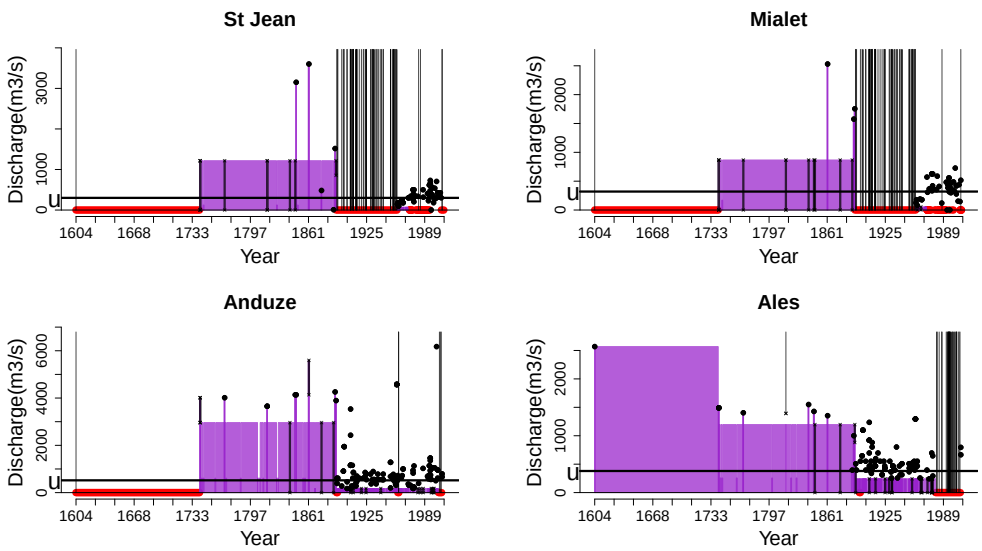

FIGURE 2. Extracted peaks-over-threshold at the four considered stations. Violet segments and areas represent data of type 2 and 3 available before declustering. Missing days are shown in red. Grey segments (*resp.* black points) are data of type 2 and 3 (*resp. 1*) belonging to an extracted multivariate cluster. The declustering threshold is represented by the horizontal black line. Vertical Grey lines are drawn at days which are missing at the considered location but which belong to a cluster, due to a threshold excess at another location.

After declustering and censoring below threshold, the data set is made of $n_{\mathbf{v}} = 125$ d-variate cluster maxima $\{\mathbf{C}^{\vee}_{t_i}, 1 \leq i \leq n_v\}$. The empirical mean cluster size is $\hat{\tau} = 1.248$, which is to be used as a normalizing constant for the number of inter-cluster days. Namely, m dependent inter-cluster observations contribute to the likelihood as $m/\hat{\tau}$ independent ones would do (see *e.g.* Beirlant et al. (2004), Chap. 10 or Coles (2001), Chap. 8). As for those inter-cluster observations, $n_{\text{bel}} = 7562$ data points are below thresholds and only 9 days are completely missing (no recording at any location). The remaining $n'_{\mathbf{v}} = 140\,674$ days are undetermined, and must be taken into account in the likelihood expression. They can be classified into 34 homogeneous temporal blocks (*i.e.* all the days within a

given block contain the same information), typically, between two recorded annual maxima. The block sizes are $n'_i (1 \leq i \leq 34)$, so that $\sum_{i=1}^{34} n'_i = n'_{\mathbf{v}}$.

Figure 3 shows bi-variate plots of the extracted cluster maxima together with undetermined blocks. Exact data are represented by points; One coordinate missing or censored yields a segment and censoring at both locations results in a rectangle. The plots show the asymmetrical nature of the problem under study: the quantity of available data varies from one pair to another (compare, *e.g.*, the number of points available respectively for the pair Saint-Jean/Mialet and Saint-Jean/Alès). Joint modeling of excesses thus appears as a way of transferring information from one location to another. Also, the most extreme observations seem to occur simultaneously (by pairs): They are more numerous in the upper right corners than near the axes, which suggests the use of a dependence structure model for asymptotically dependent data such as the Dirichlet mixture (see Section 3.2).

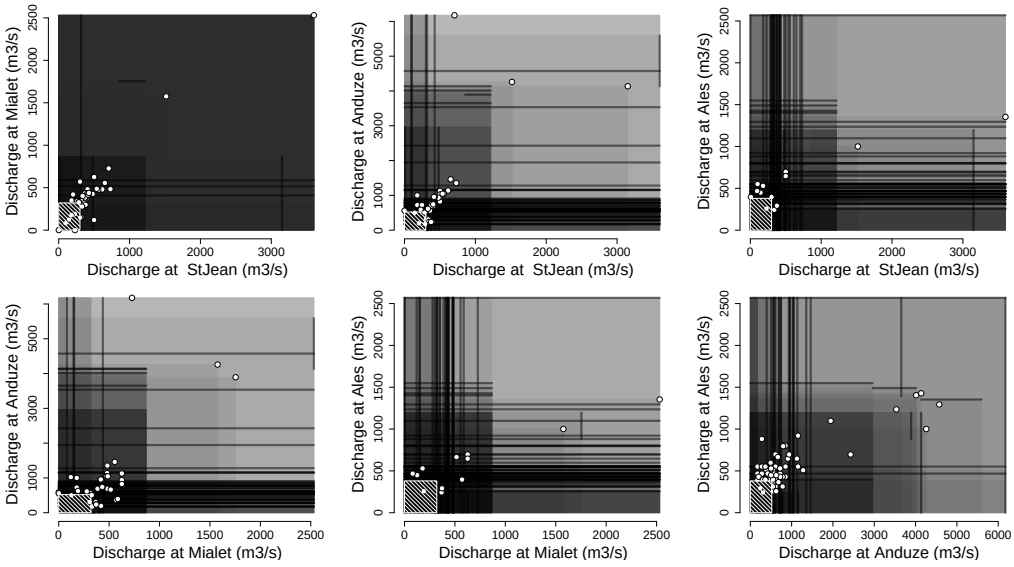

FIGURE 3. Bi-variate plots of the 124 simultaneous stream-flow records (censored cluster maxima) at the four stations, and of the 34 undetermined data blocks defined in section 2.2, over the whole period 1604-2010. Points represent exact data, gray lines and squares respectively represent data for which one (*resp.* two) coordinate(s) is (are) censored. Data superposition is represented by increased darkness. The striped rectangle at the origin is the region where all coordinates are below threshold.

3. Multivariate peaks-over-threshold model

This section provides a short description of the statistical model used for estimating the joint distribution of excesses above high thresholds. A more exhaustive statistical description is given in the above mentioned forthcoming paper. For an overview of statistical modeling of extremes in hydrology, the reader may refer *e.g.* to Katz et al. (2002). Also, Davison and Smith (1990) focus on the uni-variate case

and Coles and Tawn (1991) review the most classical multivariate extreme value models.

3.1. **Marginal model.** After declustering, the extracted cluster maxima are assumed to be independent from each other. Their margins (values of the cluster maxima at each location considered separately) can be modeled by a Generalized Pareto distribution above threshold, provided that the latter is chosen high enough (Davison and Smith, 1990; Coles, 2001). Let $Y^{\vee}_{j,t_i}$ be the (possibly unobserved) maximum water discharge at station j, in cluster i and let $F^{\mathbf{v}}_j$ the marginal cumulative distribution function (*c.d.f.*) below threshold. The marginal probability of an excess above threshold is denoted ζ_j $(1 \leq j \leq d)$. Following common practice (*e.g.* Coles and Tawn, 1991; Davison and Smith, 1990; Ledford and Tawn, 1996), ζ_j is identified with its empirical estimate $\hat{\zeta}_j$, which is obtained as the proportion of intra-cluster days (after uni-variate declustering) among the non-missing days for the considered margin and threshold. For $\mathbf{v}$ as above, it yields $\boldsymbol{\zeta} \simeq (0.0021, 0.0022, 0.0022, 0.0020)$.

The marginal models are thus

$$\begin{aligned} F_j^{(\xi_j,\sigma_j)}(y) &= \mathbf{P}(Y^{\vee}_{j,t_i} < y | \xi_j, \sigma_j) \ , \qquad (1 \leq j \leq d) \\ &= \begin{cases} 1 - \zeta_j \left(1 + \xi_j \frac{y - v_j}{\sigma_j}\right)^{-1/\xi} & (\text{if } y \geq v_j), \\ (1 - \zeta_j) F^{\mathbf{v}}_j(y) & (\text{if } y < v_j). \end{cases} \end{aligned}$$

The marginal parameters are gathered into a $2d$-dimensional vector

$$\chi = (\log(\sigma_1), \ldots, \log(\sigma_d), \xi_1, \ldots, \xi_d) \ ,$$

and the uni-variate *c.d.f.*'s are denoted by F^{χ}_j.

In a context of regional frequency analysis, it is further assumed that the shape parameter of the marginal GP distributions is identical for all catchments, i.e. $\xi_1 = \cdots = \xi_d$.

3.2. **Dependence structure.** In order to apply probabilistic results from multivariate extreme value theory, it is convenient to handle Fréchet distributed variables X_{j,t_i}, so that $P(X_{j,t_i} < x) = e^{-\frac{1}{x}}$, $x > 0$. This is achieved by defining a marginal transformation

$$\mathcal{T}^{\chi}_j(y) = -1/\log\left(F^{\chi}_j(y)\right),$$

and letting $X_{j,t_i} = \mathcal{T}^{\chi}_j(Y_{j,t_i})$. The dependence structure is then defined between the Fréchet-transformed data. One key assumption underlying multivariate extreme value models is that random vectors $\mathbf{Y}_t = (Y_{1,t}, \ldots, Y_{d,t})$ are regularly varying (see *e.g.* Resnick, 1987, 2007; Beirlant et al., 2004; Coles and Tawn, 1991). Multivariate regular variation (MRV) can be expressed as a radial homogeneity property of the distribution of the largest observations: For any region $A \subset (\mathbb{R}^+)^d$ bounded away from 0, if we denote $r.A = \{\mathbf{x} \in \mathbb{R}^d : \frac{1}{r}\mathbf{x} \in A\}$, then, for large r_0's and for $r > r_0$, MRV and transformations to unit-Fréchet imply that

$$r\,\mathbf{P}(\mathbf{X} \in r.A) \underset{r_0 \to \infty, r > r_0}{\sim} r_0\,\mathbf{P}(\mathbf{X} \in r_0.A)\,. \tag{1}$$

Switching to a pseudo-polar coordinates system, let $R = \sum_{j=1}^{d} X_j$ denote the radius and $\mathbf{W} = (\frac{X_1}{R}, \ldots, \frac{X_d}{R})$ denote the angular component of the Fréchet rescaled data. In this context, $\mathbf{W}$ is a point on the simplex $\mathbf{S}_d$: $\sum_{j=1}^{d} W_j = 1$, $W_j \geq 0$. Then (1) implies that, for any angular region $B \subset \mathbf{S}_d$,

$$\mathbf{P}(\mathbf{W} \in B \,|\, R > r_0) \underset{r_0 \to \infty}{\longrightarrow} H(B) \tag{2}$$

where H is the so-called 'angular probability measure', *i.e.* the distribution of the angles corresponding to large radii. Since in addition, $\mathbf{P}(R > r_0) \underset{r_0 \to \infty}{\sim} \frac{d}{r_0}$, the joint behavior of large excesses is entirely determined by H.

As an illustration of this notion of angular distribution, Figure 4 shows two examples of simulated bi-variate data sets, with two different angular distributions and same Pareto-distributed radii. H's density is represented by the pale red area. In the left panel, H has most of its mass near the end points of the simplex (which is, in dimension 2, the segment $[(1,0),(0,1)]$, represented in blue on Figure 4) and the extremes are weakly dependent, so that events which are large in both components are scarce. In the limit case where H is concentrated at the end-points of the simplex (not shown), the pair is said to be asymptotically independent. In contrast, the right panel shows a case of strong dependence: H is concentrated near the middle point of the simplex and extremes occur mostly simultaneously.

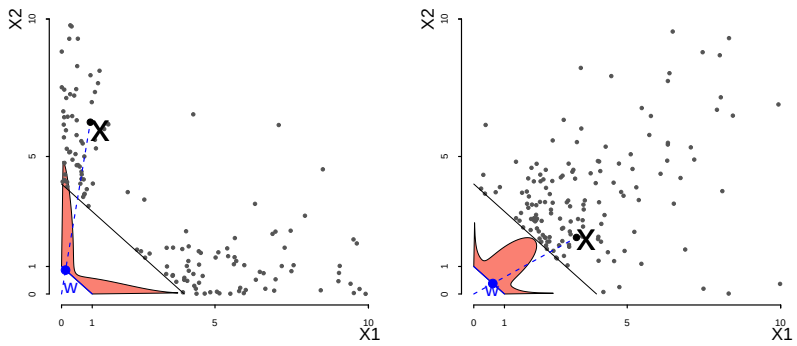

FIGURE 4. Two Examples of bivariate dependence structures of excesses above a radial threshold.
Grey points: simulated bivariate data. Pale red area: density of the angular distribution. Blue point: one randomly chosen angle $\mathbf{W}$, corresponding to the observation $\mathbf{X}$ (black point).

Contrary to the limit distribution of uni-variate excesses, H does not have to belong to any particular parametric family. The only constraint on H is due to the standard form of the X_j's: H is a valid angular distribution if and only if

$$\int_{\mathbf{S}_d} w_j \, \mathrm{d}H(\mathbf{w}) = \frac{1}{d} \quad (1 \leq j \leq d)\,. \tag{3}$$

In this paper, H is chosen in the Dirichlet mixture model (Boldi and Davison, 2007), which can approach any valid angular distribution. In short, a Dirichlet

distribution with shape $\nu \in \mathbb{R}^+$ and center of mass $\boldsymbol{\mu} \in \mathbf{S}_d$ has density

$$\mathrm{diri}_{\nu,\boldsymbol{\mu}}(\mathbf{w}) = \frac{\Gamma(\nu)}{\prod_{j=1}^d \Gamma(\nu\mu_j)} \prod_{j=1}^d w_j^{\nu\mu_j - 1} .$$

The density of a Dirichlet mixture distribution is therefore a weighted average of Dirichlet densities. A parameter for a k-mixture is thus of the form

$$\psi = \big((p_1, \ldots, p_k), (\boldsymbol{\mu}_{\cdot,1}, \ldots, \boldsymbol{\mu}_{\cdot,k}), (\nu_1, \ldots, \nu_k)\big) ,$$

with weights $p_m > 0$, $\sum_m p_m = 1$, which will be denoted by $\psi = \big(p_{1:k}, \boldsymbol{\mu}_{\cdot,1:k}, \nu_{1:k}\big)$. The corresponding mixture density is

$$h_\psi(\mathbf{w}) = \sum_{m=1}^k p_m \, \mathrm{diri}_{\nu,\boldsymbol{\mu}_{\cdot,m}}(\mathbf{w}) .$$

As for the moment constraint (3), it is satisfied if and only if

$$\sum_{m=1}^k p_m \boldsymbol{\mu}_{\cdot,m} = (1/d, \ldots, 1/d) . \tag{4}$$

In other terms, the center of mass of the $\boldsymbol{\mu}_{\cdot,1:m}$'s, with weights $p_{1:m}$, must lie at the center of the simplex.

3.3. **Estimation using censored data.** Data censorship is the main technical issue in this paper. This section exposes the matter as briefly as possible. For the sake of readability, technical details and full statistical justification have been gathered in the above mentioned unpublished paper.

In order to account for cenored data overlapping threshold and censored or missing components in the likelihood expression, it is convenient to write the model in terms of a Poisson point process, with intensity measure determined by H. More precisely, after marginal standardization, the time series of excesses above large thresholds can be described as a Poisson point process (PRM),

$$\sum_{t=1}^n \mathbb{1}_{(t,\mathbf{X}_t)} \sim \mathrm{PRM}(\,\mathrm{d}s \times \,\mathrm{d}\lambda) \quad \text{on } [0,n] \times A_{\boldsymbol{u}} ,$$

where n is the length of the observation period, $A_{\boldsymbol{u}}$ is the 'extreme' region on the Fréchet scale, $A_{\boldsymbol{u}} = [0,\infty]^d \setminus [0,u_1] \times \cdots \times [0,u_d]$, above Fréchet thresholds $u_j = \mathcal{T}_j^\chi(v_j) = -1/\log(1-\boldsymbol{\zeta}_j)$. The notation $\mathrm{d}s$ stands for the Lebesgue measure on $[0,1]$ and λ is the so-called 'exponent measure', which is related to the angular distribution's density h *via*

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\mathbf{x}}(\mathbf{x}) = d.h(\mathbf{w})r^{-(d+1)} \qquad \Big(r = \sum_{j=1}^d x_j \,,\, \mathbf{w} = \mathbf{x}/r\Big) .$$

This Poisson model has been widely used for statistical modeling of extremes (Coles, 2001; Coles and Tawn, 1991; Joe et al., 1992). The major advantage in our context is that it allows to take into account the undetermined data (which cannot be ascertained to be below nor above threshold), as they correspond to events of the kind

$$\mathbf{N}\Big\{\big[t_i', t_i' + n_i'\big] \times \Big([0,\infty]^d \setminus [0, \mathcal{T}_1^\chi(R_{1,t_i'})] \times \ldots [0, \mathcal{T}_d^\chi(R_{d,t_i'})]\Big)\Big\} = 0 ,$$

where $\mathbf{N}\{\cdot\}$ is the number of points from the Poisson process in a given region.

In our context, h is a Dirichlet mixture density: $h = h_\psi$. Let $\theta = (\chi, \psi)$ represent the parameter for the joint model, and λ_ψ be the Poisson intensity associated with h_ψ. The likelihood in the Poisson model, in the absence of censoring, is

$$\mathcal{L}_{\mathbf{v}}\left(\{\mathbf{y}_t\}_{1\leq t\leq n}, \theta\right) \propto e^{-n\,\lambda_\psi(A_{\boldsymbol{u}})} \prod_{i=1}^{n_{\mathbf{v}}} \left\{ \frac{\mathrm{d}\lambda_\psi}{\mathrm{d}\mathbf{x}}(\mathbf{x}_{t_i}) \prod_{j:y_{j,t_i}>v_j} J_j^\chi(y_{j,t_i}) \right\}. \tag{5}$$

The J_j^χ's are the Jacobian terms accounting for the transformation $\mathbf{y} \to \mathbf{x}$.

The likelihood function in presence of such undetermined data and of censored data above threshold is obtained by integration of (5) in the direction of censorship. These integrals do not have a closed form expression. In a Bayesian context, a Markov Chain Monte-Carlo (MCMC) algorithm is built in order to sample from the posterior distribution, and the censored likelihood is involved at each iteration. Rather than using numerical approximations, whose bias may be difficult to assess, one option is to use a *data augmentation* framework (see *e.g.* Tanner and Wong, 1987; Van Dyk and Meng, 2001). The main idea is to draw the missing coordinates from their full conditional distribution in a Gibbs-step of the MCMC algorithm. Again, technicalities are omitted here.

4. Results

In this section, the multivariate extreme model with Dirichlet mixture dependence structure is fitted to the data from the Gardons, including all historical data and assuming a regional shape parameter. This regional hypothesis is confirmed (not rejected) by a likelihood ratio test: the p-value of the χ^2 statistic is 0.16. To assess the added value of taking into account historical data on the one hand, and of a regional analysis on the other hand, inference is also made without the regional shape assumption and considering only the systematic measurement period (starting from January, 1892). Thus, in total, four model fits are performed.

For each of the four experiments, 6 chains of 10^6 iterations are run in parallel, which requires a moderate computation time[2]. Using parallel chains allow to check convergence using standard stationarity and mixing tests (Heidelberger and Welch (1983)'s test , Gelman and Rubin (1992)'s variance ratio test), available in the R statistical software. In the remainder of this section, all posterior predictive estimates are computed using the last $8\,10^5$ iterations of the chain obtaining the best stationarity score.

Figure 5 shows posterior histograms of the marginal parameters, together with the prior density. The posterior distributions are much more concentrated than the priors, indicating that marginal parameters are identifiable in each model. Also, the shape and scale panels are almost symmetric: a posterior distribution granting most weight to comparatively high shape parameters concentrates on comparatively low scales. This corroborates the fact that frequentist estimates of the shape and the scale parameter are negatively correlated (Ribereau et al., 2011). In the regional model as well as in the local one, the posterior variance of each parameter is reduced when taking into account historical data (except for the

[2]The execution time ranged from approximately 3h30′ to 4h30′ for each chain on a standard processor Intel 3.2 GHz.

scale parameter at Anduze, for the local model). This confirms the general fact that taking into account more data tends to reduce the uncertainty of parameter estimates.

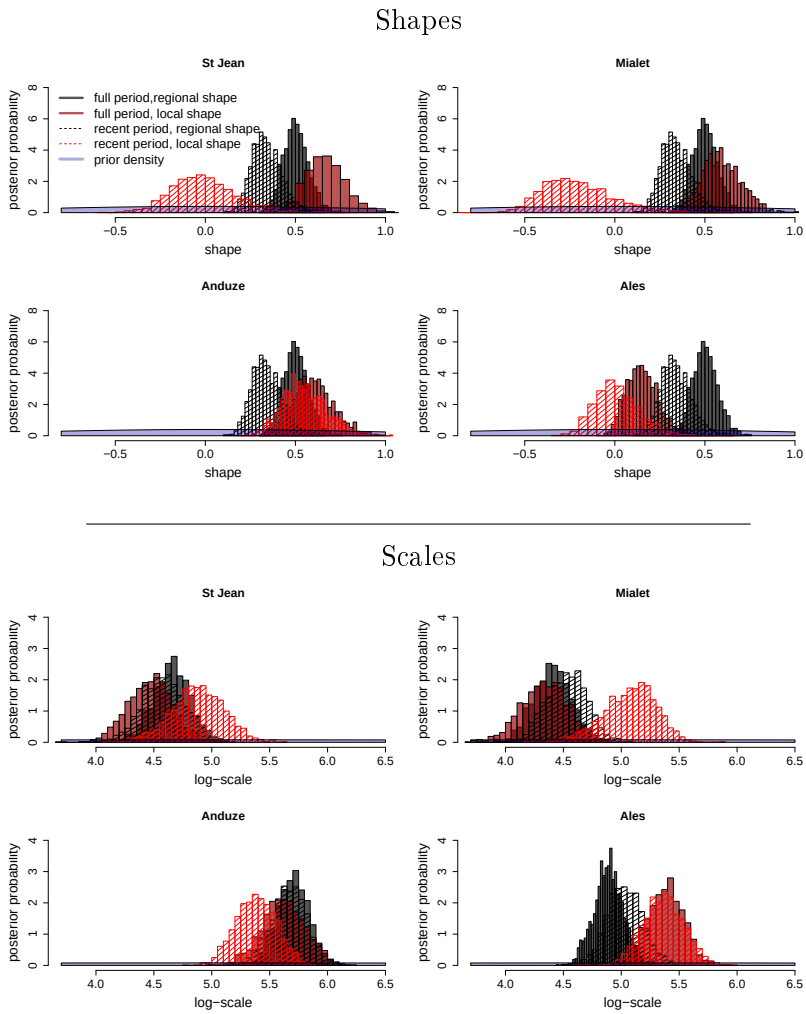

FIGURE 5. Prior and posterior distributions of the shape parameter (upper panel) and of the logarithm of the scale parameter (lower panel) at the four locations, estimated with or without historical data, in a regional framework or not.

Figure 6 shows posterior mean estimates of the return levels at each location, together with credible intervals based on posterior $0.05 - 0.95$ quantiles, in the four inferential frameworks. The return levels appear to be very sensitive to model choice: overall, taking into account the whole period increases the estimated return levels. In terms of mean estimate, the effect of imposing a global shape parameter varies from one station to another, as expected. For those return levels, the posterior credibility intervals seem to depend more on the mean return levels than on the choice of a regional or local framework. This seems at odds with the previous findings of reduced intervals for marginal parameters. However, one must

note that the width of return level credibility intervals depends not only on that of the parameters, but also on the value of the mean estimates. Larger parameter estimates involve larger uncertainty in terms of return levels.

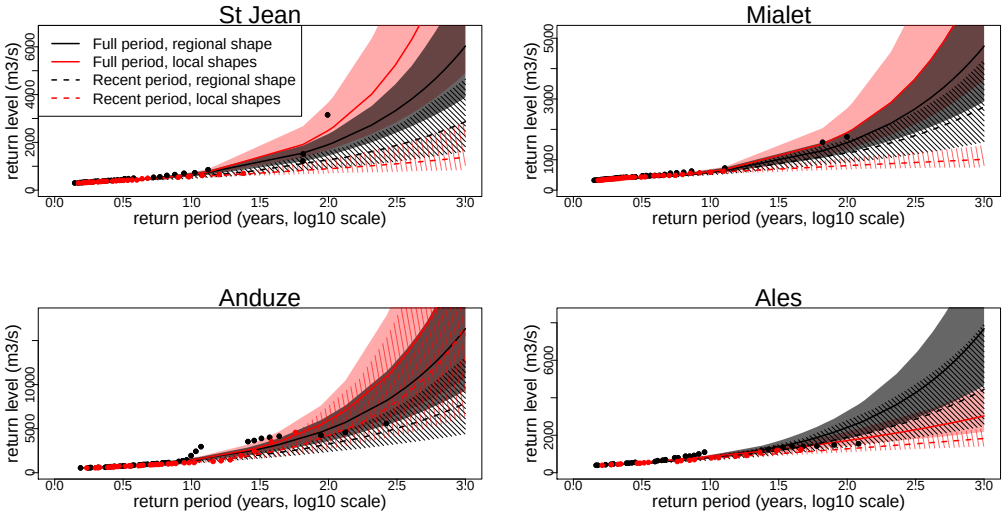

FIGURE 6. Return level plots at each location using four inferential frameworks with 90% posterior quantiles. Dotted lines and hatched areas: data from he recent period only; Solid lines and shaded area: Full data set; Black lines and Grey area: Regional analysis, global shape parameter; Red lines and shaded red area: local shape parameters. Black (*resp.* Red) points: observed data plotted at the corresponding empirical return period using the whole (*resp.* recent) data set.

In addition to uni-variate quantities of interest such as marginal parameters or return level curves, having estimated the dependence structure gives access to multivariate quantities. Figure 7 shows the posterior mean estimates of the angular density. Since the four-variate version of the angular distribution cannot be easily represented, the bivariate marginal versions of the angular distribution are displayed instead. Here, the unit simplex (which was the diagonal blue segment in Figure 4) is represented by the horizontal axis, so that H is a distribution function on $[0, 1]$. As could be expected in view of Figure 3, extremes are rather strongly dependent. Moreover, the posterior distribution is overall well concentrated around the mean estimate.

The predictive angular distribution allows to estimate conditional probabilities of exceedance of high thresholds. As an example, figure 8 displays, for the six pairs $1 \leq j < i \leq 4$, the posterior estimates of the conditional tail distribution functions $P(Y_i^{\vee} > y | Y_j^{\vee} > v_j)$ at location i, conditioned upon an excess of the threshold v_j at another location j. The predictive tail functions in the DM model concur with the empirical estimates for moderate values of y. For larger values, the empirical error grows and no empirical estimate exists outside the observed

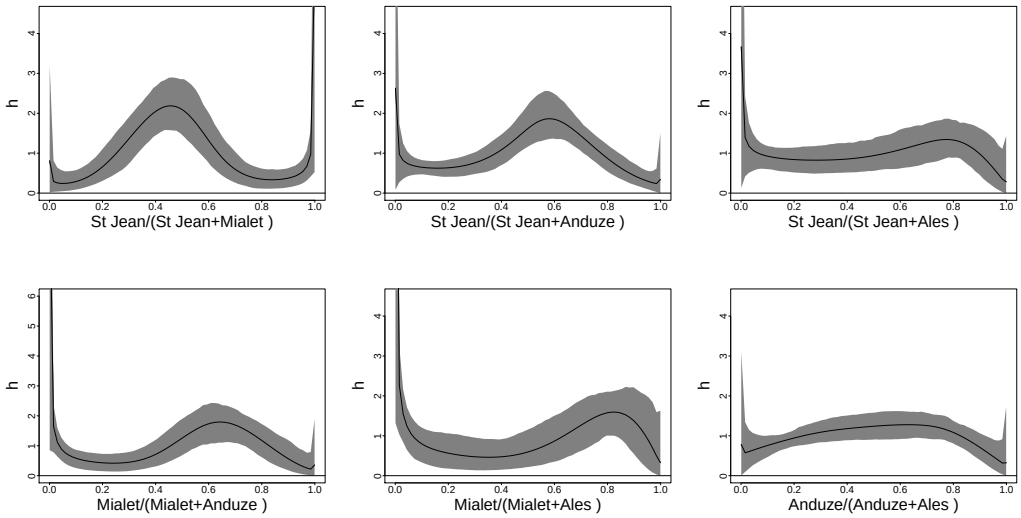

FIGURE 7. Posterior predictive bi-variate angular densities (black lines) with posterior $0.05 - 0.95$ quantiles (Grey areas).

domain. However, the DM estimates are still defined and the size of the error region remains comparatively small.

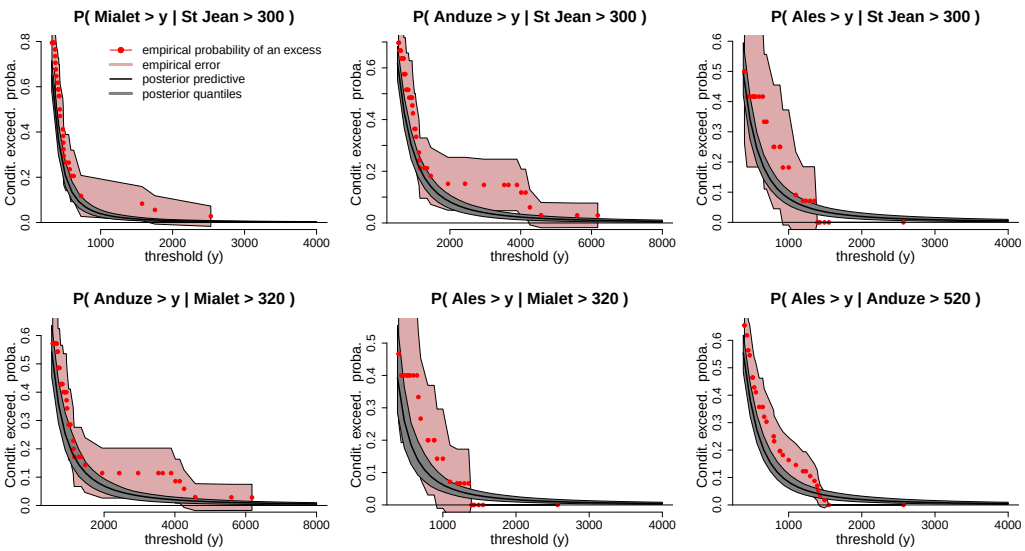

FIGURE 8. Conditional tail distributions. Black line and Grey area: posterior mean estimate and posterior 90% credible intervals (posterior quantiles); red points: empirical tail function computed at the recorded points above threshold; pale red area: 90% Gaussian confidence intervals around the empirical estimates.

Finally, one commonly used measure of dependence at asymptotically high levels between pairs of locations is defined by (Coles et al., 1999):

$$\boldsymbol{\chi}_{i,j} = \lim_{x\to\infty} \frac{P(X_i > x, X_j > x)}{P(X_j > x)} = \lim_{x\to\infty} P(X_i > x \,|X_j > x\,)\,,$$

where X_i, X_j are the Fréchet-transformed variables at locations i and j. Since X_i and X_j are identically distributed, $\boldsymbol{\chi}_{i,j} = \boldsymbol{\chi}_{j,i}$. From its definition, $\boldsymbol{\chi}_{i,j}$ is comprised between 0 and 1; small values indicate weak dependence at high levels whereas values close to 1 are characteristic of strong dependence. In the extreme case $\boldsymbol{\chi} = 0$, the variables are asymptotically independent. In the case of Dirichlet mixtures, $\boldsymbol{\chi}_{i,j}$ has an explicit expression formed of incomplete Beta functions (Boldi and Davison, 2007, eq. (9)). Figure 9 shows posterior box-plots of $\boldsymbol{\chi}$ for the six pairs. The strength of the dependence and the amount of uncertainty varies from one pair to another, but mean estimates are overall large (greater than 0.4), indicating strong asymptotic dependence.

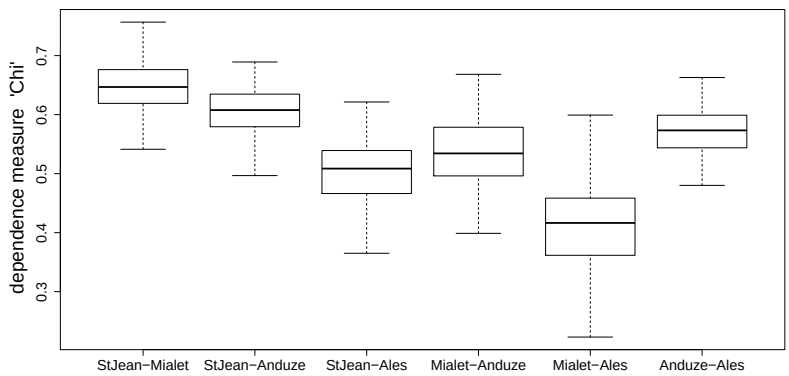

FIGURE 9. Dependence measure $\boldsymbol{\chi}_{i,j}$ for the six pairs of locations: Posterior box-plot.

In order to verify the consistency of those results with observed data, empirical quantities $P(X_i > x|X_j > x)$ have been computed and are displayed in Figure 10. More precisely, it is easy to see that

$$P(X_i > x|X_j > x) = P(Y_i^{\vee} > (F_i^{\chi})^{-1} \circ F_j^{\chi}(y)|Y_j^{\vee} > y)\,,$$

where F_i^{χ}, F_j^{χ} are the marginal *cdf* for location i and j, and the $Y_j^{\vee}$, $Y_i^{\vee}$'s are the observed data (cluster maxima). In Figure 10, the conditioning thresholds y are the observed values of the conditioning variable $Y_j^{\vee}$ above the initial threshold v_j, of which the estimated return period (abscissa of the red points) is taken as its mean estimate using the marginal parameter components of the posterior sample. For each such y, $(F_i^{\chi})^{-1} \circ F_j^{\chi}(y)$ is estimated by its posterior mean value, again computed from the marginal posterior sample. Then, the conditional probability of an excess by $Y_i^{\vee}$ (Y-axis value of the red points) is computed empirically. In theory, as the return period increases, the red points should come closer to the horizontal black line, which is the mean estimate of $\boldsymbol{\chi}$ computed in the Dirichlet mixture dependence model, as in Figure 9. Note that in the Dirichlet model, the limiting value $\boldsymbol{\chi}$ is already reached at finite levels because the conditional probability of an

excess on the Fréchet scale, $P(X_i > x | X_j > x)$, is constant in x. On the contrary, in an asymptotically independent model, the conditional exceeedance probability whould be decreasing towards zero. Results in Figure 10 are comforting: the mean values of $\boldsymbol{\chi}$ obtained from the Dirichet model are within the error regions of the empirical estimates. The latter are very large, compared to the posterior quantiles from the Dirichlet mixture, which illustrates the usefulness of an extreme value model for computing conditional probabilities of an excess.

This result has implications for computing the return periods of joint excesses of high thresholds. Consider, for example, the 10 years marginal return levels at two stations, (q_1, q_2). If the excesses above these threhsolds were assumed to be independent, taking into account short term temporal dependence (the mean cluster size is $\tau = 1.248$), the return period for the joint excess $(Y_1^{\vee} > q_1, Y_2^{\vee} > q_2)$ would be $10^2 * (365/\tau) = 29247.8$ years. On the contrary, accounting for spatial dependence, for example between the two first stations (St Jean and Mialet), yields an estimated return period for a joint excess of $10/\hat{\boldsymbol{\chi}}_{1,2} = 10/0.645 = 15.5$ years.

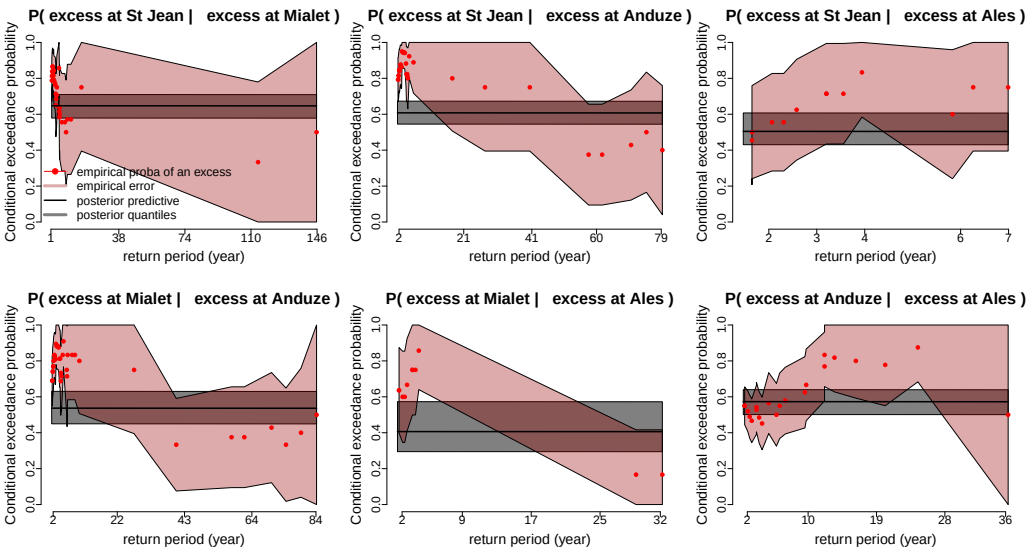

FIGURE 10. Observed conditional probability of exceedance of equally scaled thresholds. Red points: empirical estimates of conditional excesses; pale red regions, empirical standard error; horizontal black line and gray area, posterior mean and $0.05 - 0.95$ quantiles of the theoretical value in the DM model.

5. DISCUSSION

This section lists the limitations of the model used in this paper and discusses directions for improvement.

5.1. **Impact of systematic rating curve errors.** The use of historical data allows extending the period of record and hence the availability of extreme flood events. However, historical data are also usually much more uncertain than recent

systematic data, for two reasons: (i) the precision of historical water stages is limited; (ii) the transformation of these stage values into discharge values is generally based on a rating curve derived using a hydraulic model, which may induce large systematic errors.

The model used in the present paper ignores systematic errors (ii). This is because we focused on multivariate aspects through the use of the DM model to describe intersite dependence. However, systematic errors may have a non-negligible impact on marginal quantile estimates, as discussed by Neppel et al. (2010). Moreover, in a multivariate context, the impact of systematic errors on the estimation of the dependence structure is unclear at this stage and requires further evaluation. Future work will therefore aim at incorporating an explicit treatment of systematic errors, using models such as those discussed by Reis and Stedinger (2005) or Neppel et al. (2010).

5.2. **Comparing several models for intersite dependence.** The DM model used in this paper to describe intersite dependence is a valid dependence model according to multivariate extreme value theory (MEVT). Many alternative approaches, not necessarily MEVT-compatible, have been proposed in the hydrological literature on regional estimation methods. Such approaches include simply ignoring dependence (e.g. Dalrymple, 1960), the concept of 'equivalent number of sites' (Reed et al., 1999) or the use of copulas (e.g. Renard, 2011). This raises the question of the influence of the approach used to describe dependence on the following estimates:

- Marginal estimates, typically quantile estimates at each site. While the impact of ignoring dependence altogether has been studied by several authors (Stedinger, 1983; Hosking and Wallis, 1988; Madsen and Rosbjerg, 1997; Renard and Lang, 2007), the impact of alternative dependence models is less clear. In particular, since marginal estimates do not directly use the dependence model, it remains to be established whether or not different dependence models (e.g. asymptotically dependent vs. asymptotically independent) yield significantly different results.
- Joint or conditional estimates, as illustrated in Figures 3, 8 and 10 for instance. The dependence model obviously plays a much more important role in this case.

Such comparison has not been attempted in this paper because the use of censored historical data makes the application of standard methods like copulas much more challenging.

5.3. **The treatment of intersite dependence in a highly dimensional context.** As illustrated in the case study, the DM model is applicable in moderate dimension d=4. However, such semi-parametric approach is not geared toward highly-dimensional contexts (e.g. spatial rainfall using dozens or hundreds of rain gauges, or gridded data sets). Practical approaches for highly-dimensional multivariate extremes have been mostly proposed in the context of block maxima, using the theory of max-stable processes (De Haan, 1984; Smith, 1990; Schlather, 2002; Westra and Sisson, 2011). Estimation procedures *e.g.* using composite likelihood methods exist for such processes (Padoan et al., 2010), along with descriptive tools *e.g.* to define and estimate extremal dependence coefficients such as the madogram

(Cooley et al., 2006). However, the development of models adapted to peaks-over-threshold is still an area of active research in a highly-dimensional spatial context and full modeling (which would *e.g* allow simulation of joint excesses) remain elusive. Recent theoretical advances (Ferreira and de Haan, 2012; Dombry and Ribatet, 2013) give cause to hope for, and expect, future development of spatial peaks-over-threshold models.

6. Conclusion

This paper illustrates the use of a multivariate peaks-over-threshold model to combine regional estimation and historical floods. This model is based on a semi-parametric Dirichlet Mixture to describe intersite dependence, while Generalized Pareto distributions are used for margins. A data augmentation scheme is used to enable the inclusion of censored historical flood data. The model is applied to four catchments in Southern France where historical flood data are available.

The first objective of this case study was to assess the relative impact of regional and historical information on marginal quantile estimates at each site. The main results can be summarized as follows:

- Over the four considered versions of the model, the version ignoring historical floods and performing local estimation yields estimates that may strongly differ from the other versions. The three other versions (which either use historical floods or perform regional estimation or both) yield more consistent estimates. This illustrates the benefit of extending the at-site sample using either historical or regional information, or both.
- Compared with the most complete version of the model (which enables both historical floods and regional estimation), the version only implementing regional estimation (but ignoring historical floods) yields smaller estimates of the shape parameter, and hence smaller quantiles. This result is likely specific to this particular data set, for which many large floods have been recorded during the historical period.
- Compared with the most complete version of the model, the version using historical floods but implementing local estimation yields higher quantiles for three catchments but lower quantiles on the fourth.
- The uncertainty in parameter estimates generally decreases when more information (regional, historical or both) is included in the inference. However, this does not necessarily result in smaller uncertainty in quantile estimates. This is because this uncertainty does not only depends on the uncertainty in parameter estimates, but also on the value taken by the parameters. In particular, a precise but large shape parameter may result in more uncertain quantiles than a more imprecise but lower shape parameter.

The second objective was to investigate the nature of asymptotic dependence in this flood data set, by taking advantage of the existence of extremely high joint exceedances in the historical data. Results in terms of predictive angular density suggest the existence of such dependence between every pairs of catchments of asymmetrical nature: some pairs are more dependent than others at asymptotic levels. In addition, the Dirichlet Mixture model allows to compute bi-variate conditional probabilities of large threshold exceedances, which are poorly estimated

with empirical methods. The limiting values of the conditional probabilities, theoretically obtained with increasing thresholds, are substantially non zero (they range between 0.4 and 0.65), which confirms the strength and the asymmetry of pairwise asymptotic dependence for this data set and induces multivariate return periods much shorter than they would be in the asymptotically independent case.

References

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004). *Statistics of extremes: Theory and applications*. John Wiley & Sons: New York.

Boldi, M.-O. and Davison, A. C. (2007). A mixture model for multivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):217–229.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer Verlag.

Coles, S., Heffernan, J., and Tawn, J. A. (1999). Dependence measures for extreme value analyses. *Extremes*, 2:339–365.

Coles, S. and Tawn, J. (1991). Modeling extreme multivariate events. *JR Statist. Soc. B*, 53:377–392.

Cooley, D., Naveau, P., and Poncet, P. (2006). Variograms for spatial max-stable random fields. In *Dependence in probability and statistics*, pages 373–390. Springer.

Dalrymple, T. (1960). Flood frequency analyses. *Water-supply paper 1543-A*.

Davison, A. and Smith, R. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 393–442.

De Haan, L. (1984). A spectral representation for max-stable processes. *The annals of probability*, pages 1194–1204.

De Haan, L. and De Ronde, J. (1998). Sea and wind: Multivariate extremes at work. *Extremes*, 1:7–45.

Dombry, C. and Ribatet, M. (2013). Functional regular variations, pareto processes and peaks over threshold.

Ferreira, A. and de Haan, L. (2012). The generalized pareto process; with a view towards application and simulation. *arXiv preprint arXiv:1203.2551v2*.

Ferro, C. and Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):545–556.

Gaume, E., Gaal, L., Viglione, A., Szolgay, J., Kohnova, S., and Bloschl, G. (2010). Bayesian mcmc approach to regional flood frequency analyses involving extraordinary flood events at ungauged sites. *Journal of Hydrology*, 394:101–117.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

Gumbel, E. (1960). Distributions des valeurs extrêmes en plusieurs dimensions. *Publ. Inst. Statist. Univ. Paris*, 9:171–173.

Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, pages 1109–1144.

Hosking, J. (1985). Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution. *Applied Statistics*, 34:301–310.

Hosking, J. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3):339–349.

Hosking, J. and Wallis, J. R. (1988). The effect of intersite dependence on regional flood frequency analysis. *Water Resources Research*, 24:588–600.

Hosking, J. and Wallis, J. R. (1997). *Regional Frequency Analysis: an approach based on L-Moments*. Cambridge University Press, Cambridge, UK.

Jin, M. and Stedinger, J. R. (1989). Flood frequency analysis with regional and historical information. *Water Resources Research*, 25(5):925–936.

Joe, H., Smith, R. L., and Weissman, I. (1992). Bivariate threshold methods for extremes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 171–183.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in water resources*, 25(8):1287–1304.

Lang, M., Ouarda, T., and Bobee, B. (1999). Towards operational guidelines for over-threshold modeling. *Journal of Hydrology*, 225:103–117.

Leadbetter, M. (1983). Extremes and local dependence in stationary sequences. *Probability Theory and Related Fields*, 65(2):291–306.

Ledford, A. and Tawn, J. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Madsen, H., Pearson, C. P., and Rosbjerg, D. (1997a). Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events .2. regional modeling. *Water Resources Research*, 33(4):759–769.

Madsen, H., Rasmussen, P. F., and Rosbjerg, D. (1997b). Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events .1. at-site modeling. *Water Resources Research*, 33(4):747–757.

Madsen, H. and Rosbjerg, D. (1997). The partial duration series method in regional index-flood modeling. *Water Resources Research*, 33(4):737–746.

Nadarajah, S. (2001). Multivariate declustering techniques. *Environmetrics*, 12(4):357–365.

Naulet, R., Lang, M., Ouarda, T. B., Coeur, D., Bobée, B., Recking, A., and Moussay, D. (2005). Flood frequency analysis on the ardèche river using french documentary sources from the last two centuries. *Journal of Hydrology*, 313(1):58–78.

Neppel, L., Renard, B., Lang, M., Ayral, P., Coeur, D., Gaume, E., Jacob, N., Payrastre, O., Pobanz, K., and Vinet, F. (2010). Flood frequency analysis using historical data: accounting for random and systematic errors. *Hydrological Sciences Journal–Journal des Sciences Hydrologiques*, 55(2):192–208.

O'Connel, D., Ostenaa, D., Levish, D., and Klinger, R. (2002). Bayesian flood frequency analysis with paleohydrologic bound data. *Water Resources Research*, 38(5).

Padoan, S. A., Ribatet, M., and Sisson, S. A. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105(489).

Parent, E. and Bernier, J. (2003). Bayesian pot modeling for historical data. *Journal of hydrology*, 274:95–108.

Payrastre, O., Gaume, E., and Andrieu, H. (2011). Usefulness of historical information for flood frequency analyses: Developments based on a case study. *Water Resources Research*, 47.

Reed, D. W., Faulkner, D. S., and Stewart, E. J. (1999). The forgex method of rainfall growth estimation - ii: Description. *Hydrology and Earth System*

Sciences, 3(2):197–203.

Reis, D. and Stedinger, J. R. (2005). Bayesian mcmc flood frequency analysis with historical information. *Journal of Hydrology*, 313(1-2):97–116.

Renard, B. (2011). A bayesian hierarchical approach to regional frequency analysis. *Water Resources Research*, 47.

Renard, B. and Lang, M. (2007). Use of a gaussian copula for multivariate extreme value analysis: some case studies in hydrology. *Advances in Water Resources*, 30(4):897–912.

Resnick, S. (1987). *Extreme values, regular variation, and point processes, volume 4 of Applied Probability. A Series of the Applied Probability Trust.* Springer-Verlag, New York.

Resnick, S. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling.* Springer Series in Operations Research and Financial Engineering.

Ribereau, P., Naveau, P., and Guillou, A. (2011). A note of caution when interpreting parameters of the distribution of excesses. *Advances in Water Resources*, 34(10):1215–1221.

Sabourin, A. and Naveau, P. (2013). Bayesian dirichlet mixture model for multivariate extremes: A re-parametrization. *Computational Statistics & Data Analysis*, DOI 10.1016/j.csda.2013.04.021.

Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes*, 5(1):33–44.

Smith, R. (1994). Multivariate threshold methods. *Extreme Value Theory and Applications*, 1:225–248.

Smith, R., Tawn, J., and Coles, S. (1997). Markov chain models for threshold exceedances. *Biometrika*, 84(2):249–268.

Smith, R. L. (1990). Max-stable processes and spatial extremes. *Unpublished manuscript, Univer.*

Stedinger, J. R. (1983). Estimating a regional flood frequency distribution. *Water Resources Research*, 19:503–510.

Stedinger, J. R. and Cohn, T. A. (1986). Flood frequency-analysis with historical and paleoflood information. *Water Resources Research*, 22(5):785–793.

Stephenson, A. (2003). Simulating multivariate extreme value distributions of logistic type. *Extremes*, 6(1):49–59.

Stephenson, A. (2009). High-dimensional parametric modelling of multivariate extreme events. *Australian & New Zealand Journal of Statistics*, 51(1):77–88.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Tasker, G. D. and Stedinger, J. R. (1987). Regional regression of flood characteristics employing historical information. *Journal of Hydrology*, 96:255–264.

Tasker, G. D. and Stedinger, J. R. (1989). An operational gls model for hydrologic regression. *Journal of Hydrology*, 111:361:375.

Van Dyk, D. and Meng, X. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50.

Westra, S. and Sisson, S. A. (2011). Detection of non-stationarity in precipitation extremes using a max-stable process model. *Journal of Hydrology*, 406(1):119–128.

(Anne Sabourin) Laboratoire des Sciences du Climat et de l'Environnement, CNRS-CEA-UVSQ, 91191 Gif-sur-Yvette, France or
Universite de Lyon, CNRS UMR 5208,Universite de Lyon 1, Institut Camille Jordan ,43 blvd. du 11 novembre 1918, F-69622 Villeurbanne cedex , France
E-mail address: anne.sabourin@lsce.ipsl.fr

(Benjamin Renard) Institut National de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture, Centre de Lyon, 5 rue de la Doua - CS70077, 69626 VILLEURBANNE Cedex, France

Chapitre 5

Conclusion

5.1 Résumé de la thèse et discussion

Cette thèse est consacrée au développement de méthodes d'inférence bayésienne pour les extrêmes multivariés, dans des modèles aussi élargis que possible, c'est à dire, faisant peu d'hypothèses a priori sur la structure de dépendance asymptotique. Dans un premier temps, une méthode a été proposée pour prendre en compte les estimations issues de plusieurs modèles d'extrêmes existants, sans nécessairement avoir à choisir un seul modèle. En utilisant le cadre du *Bayesian model averaging*, les différentes mesures angulaires prédictives a posteriori peuvent être moyennées en fonction du poids a posteriori de chaque modèle. L'estimateur obtenu correspond à l'estimateur de moyenne a posteriori dans l'union disjointe des modèles moyennés. Le support de la loi a posteriori est plus large que dans chacun des modèles, ce qui permet potentiellement de mieux représenter l'incertitude. En pratique, on peut comparer la performance des estimateurs à l'aide de règles de scores (*scoring rules*, Gneiting et Raftery (2007)). On constate sur des données simulées que les estimateurs obtenus avec le BMA obtiennent en moyenne de meilleurs scores que les estimateurs obtenus par sélection de modèle. Cependant, lorsque la taille de l'échantillon augmente, la loi a posteriori se concentre autour de la vraie distribution, ou, dans le cas (fréquent) où le modèle est mal spécifié (c'est à dire, lorsque la vraie loi des observations n'appartient à aucun des modèles moyennés), dans les régions du « modèle union » à plus faible distance (au sens de la divergence de Kullback-Leibler) de la vraie distribution. Lorsque cette région est réduite à un point, ceci revient à « sélectionner » le modèle qui le contient. Le BMA implémenté avec un nombre fini de modèles ne permet donc pas de produire des estimateurs approchant n'importe quelle mesure angulaire.

Cette limitation peut être contournée par l'utilisation d'un modèle de mélange de Dirichlet dont le nombre maximal de composants est élevé. En théorie, ceci revient à implémenter le BMA avec un grand nombre de sous-modèles, chacun étant un mélange à nombre fixé de composants. En pra-

tique, l'inférence a lieu de manière très différente : au lieu d'estimer la loi a posteriori dans chaque sous-modèle séparément, on utilise un algorithme permettant d'échantillonner directement la loi a posteriori dans le modèle union, en effectuant des « sauts » réversibles d'un sous-modèle à l'autre, d'où le nom de *reversible-jump algorithm*. La difficulté majeure rencontrée avec ce modèle et l'algorithme initialement proposés par Boldi et Davison (2007), tenait à la nécessité de construire des paramètres de mélange vérifiant une contrainte barycentrique, pour que la mesure angulaire satisfasse la contrainte de moments (1.16). Dans la deuxième partie de cette thèse, la contrainte sur les paramètres est levée par une re-paramétrisation du modèle. Sous cette nouvelle forme, la contrainte de moments est automatiquement satisfaite et chaque sous-modèle (à nombre fixé de composants) est un espace produit. Ceci permet de construire un algorithme à sauts réversibles dont les « sauts » sont définis simplement par l'ajout ou le retrait d'un composant, et dont les mouvements « intra modèle » peuvent être définis de manière flexible, en fonction des données. Sur le plan théorique, le caractère ϕ-irréductible de la chaîne de Markov correspondant à l'algorithme ainsi construit est établi. Les expériences menées sur des données simulées et réelles montrent que les propriétés de « mixing » de la chaîne de Markov sont considérablement améliorées par rapport à la version originale de l'algorithme. Concernant les propriétés asymptotiques du modèle, la consistance faible de la loi a posteriori est prouvée pour les cas où la vraie distribution appartient à l'adhérence de Kullback du modèle. Le temps a manqué pour délimiter l'étendue de cette adhérence. Il semble possible de montrer que les mesures angulaires à densité h bornée, ou telles que $\int_{\boldsymbol{S}_d} h \log(h) < \infty$, en font partie, en utilisant par exemple une approximation uniforme par des polynômes de Bernstein sur le simplexe (Lorentz, 1953, chapitre 2.9) et en généralisant la preuve de Petrone et Wasserman (2002) au cas multivarié. La question reste ouverte pour les mesures angulaires non dominées, ou à distance de Kullback infinie de la mesure de Lebesgue.

Le troisième volet de cette thèse consiste à utiliser le modèle de mélange de Dirichlet re-paramétré pour estimer la loi jointe des crues dans la région des Gardons, dans le sud de la France, en utilisant des crues historiques. Les données anciennes sont, pour la plupart, censurées, et le seuil de censure est variable dans le temps. Cette variabilité du seuil, ainsi que l'incertitude sur le caractère « extrême » ou non de certaines données, induite par la censure, sont prises en compte dans le cadre d'un modèle de Poisson sur les régions « loin de l'origine ». Une des conséquences de la censure est l'absence d'expression explicite pour la vraisemblance, qui s'écrit comme une intégrale, sur des régions rectangulaires, de la vraisemblance classique d'un processus de Poisson. Cette difficulté de taille pour l'implémentation d'un algorithme MCMC est traitée par « augmentation de données », c'est-à-dire en introduisant des variables d'augmentation qui « remplacent » les données manquantes ou censurées et dont on connaît la loi conditionnellement aux autres données et aux paramètres. Ces données virtuelles sont traitées comme des paramètres

supplémentaires à échantillonner. Ceci revient à augmenter la dimension de l'espace des paramètres et ralentit la convergence de l'algorithme, mais les temps d'exécution pour des données de dimension modérée (quatre ou cinq) restent raisonnables (quelques heures sur un ordinateur de bureau). Le modèle est validé sur des données simulées, puis censurées d'après un schéma semblable au schéma de censure des Gardons.

Un des objectifs de cette modélisation, d'un point de vue hydrologique, était d'améliorer l'estimation des paramètres des lois marginales par la prise en compte des données plus anciennes, et mesurées sur l'ensemble des sites de la région, en faisant l'hypothèse d'un paramètre de forme pour les lois de Pareto constant à l'échelle régionale. En effet, un dépassement de seuil enregistré sur une des stations, associé à une structure de dépendance, apporte une information sur l'intensité d'un éventuel dépassement de seuil aux sites voisins. L'ajustement du modèle sur les données des Gardons montre que les estimations de niveaux de retour (quantiles extrêmes) varient considérablement selon que l'on prend en compte ou non les données historiques censurées. En terme de marges d'incertitude, la distribution a posteriori des paramètres marginaux se concentre avec l'inclusion des données historiques. Celle des niveaux de retour s'élargit pour les périodes de retour les plus élevées, ce qui peut s'expliquer par le fait que l'écart type du niveau de retour ne dépend pas que de celui des paramètres marginaux : l'incertitude augmente avec la valeur du niveau estimé.

Concernant la structure de dépendance, les résultats indiquent des niveaux élevés de dépendance asymptotique entre les paires de stations, en terme de mesure de dépendance, définie par la probabilité conditionnelle de dépassement de seuils identiques sur l'échelle de Fréchet, $\boldsymbol{\chi}_{i,j} = \lim_{x\to\infty} P(X_i > x | X_j > x)$. Les estimateurs de $\boldsymbol{\chi}$ obtenus par moyenne a posteriori dans le modèle de mélange de Dirichlet varient entre 0.4 et 0.6, sur une échelle de 0 (indépendance asymptotique) à 1 (dépendance totale), selon la paire de stations considérée. Au vu de ces résultats, il n'a pas été jugé nécessaire de mener des tests d'indépendance asymptotique (Hüsler et Li, 2009; Tsai *et al.*, 2013). Cependant, sur des jeux de données dont la dépendance asymptotique serait moins évidente, la présence de censures empêcherait d'appliquer tels quels les tests existants. La construction de tests adaptés à des données censurées reste à faire. Concernant la question de l'indépendance asymptotique, dans toute cette thèse, on s'est intéressé aux modèles multivariés dans lesquels il y a dépendance asymptotique entre toutes les variables, de sorte que la totalité de la masse angulaire est portée par l'intérieur du simplexe. Un premier pas vers l'affaiblissement de cette hypothèse serait d'autoriser une concentration de la masse sur les faces du simplexe, ce pour quoi les propriétés de marginalisation des lois de Dirichlet devraient pouvoir être exploitées.

Enfin, outre l'amélioration de la connaissance statistique des caractéristiques du bassin versant, une application possible du modèle de Dirichlet concerne la simulation d'événements extrêmes sur de petites échelles spa-

tiales, et pourrait permettre par exemple d'inclure un module « événement orageux » dans un simulateur de pluies. Cette piste n'a pas été explorée, mais le fait que les lois conditionnelles et marginales dans un modèle de Dirichlet se calculent facilement devrait faciliter la tâche.

Trois packages R ont été développés, l'un pour l'implémentation du BMA, l'autre pour l'algorithme à sauts réversibles échantillonnant la loi a posteriori dans le mélange de Dirichlet, un troisième pour l'adaptation du mélange de Dirichlet aux données censurées. Ils doivent être soumis aux dépôts CRAN, éventuellement sous la forme d'un package commun.

5.2 Perspectives

Une attention particulière a été accordée au modèle de mélange de Dirichlet et une perspective prometteuse ouverte par ce modèle concerne l'extension à la grande dimension et aux processus extrêmes. Pour les applications faisant intervenir des données de grande dimension (les points d'une grille d'un modèle climatique par exemple et les problèmes d'extrêmes spatiaux en général), le modèle de Dirichlet, dans sa forme présente, n'est pas adapté. Le « Fléau de la dimension » induirait vraisemblablement une grande imprécision (un manque de concentration de la loi a posteriori) et ralentirait dramatiquement l'algorithme d'échantillonnage de type Gibbs. Une question naturelle : est-il toutefois possible d'utiliser le cadre des mélanges de Dirichlet pour des espaces de grande dimension, tout en restreignant l'espace des paramètres à titre de compensation ? J'envisage, si les conditions le permettent, de poursuivre mes travaux dans cette direction.

La recherche d'une telle généralisation spatiale est extrêmement tentante, car les lois de Dirichlet ne sont que des cas particuliers, en dimension finie, des processus de Dirichlet. Ces derniers peuvent par exemple être définis sur un espace compact, typiquement une région S, fermée, bornée dans le plan, ou un ensemble dénombrable (les points d'une grille). En toute généralité, soit $(S, \mathcal{S})$ un espace mesuré, supposé CMS. L'espace $\mathcal{P}(S)$ des lois de probabilité sur S, muni de la topologie de la convergence faible, est polonais et sa tribu borélienne est engendrée par les « projections » fini-dimensionnelles $P \mapsto (P(A_1), \ldots, P(A_k))$. Un processus de Dirichlet DP_α, dont le « paramètre » est la mesure (finie sur S) $\alpha(\,\cdot\,)$, est alors une mesure de probabilité aléatoire, définie par ses lois fini-dimensionnelles. Plus précisément, on dit qu'une mesure aléatoire $\boldsymbol{P}$, définie sur S, est distribuée selon un processus de Dirichlet DP_α si, pour toute partition $S_1, \ldots, S_d$ de S, le vecteur $(\boldsymbol{P}(S_1), \ldots, \boldsymbol{P}(S_d))$ suit une loi de Dirichlet de paramètre $(\alpha(S_1), \ldots, \alpha(S_d))$ (voir par exemple Ghosh et Ramamoorthi, 2003, chapitre 3).

Les processus de Dirichlet sont très communément employés en statistique bayésienne non paramétrique, en tant que prior sur $\mathcal{P}(S)$. Ici, ils pourraient intervenir en tant que loi des observations. En effet, on peut représenter la

distribution d'une certaine quantité d'intérêt sur une région de l'espace, par exemple le cumul de précipitations journalières, par une mesure aléatoire. On peut alors s'intéresser à la *répartition* des précipitations sur une telle région, conditionnellement à un dépassement d'un seuil élevé par une fonctionnelle adaptée, par exemple la quantité de précipitations intégrées sur la région. Cette répartition aléatoire conditionnelle peut alors être modélisée par un processus, ou un mélange de processus de Dirichlet.

Ceci contribuerait au développement de modèles d'extrêmes définis par des dépassements simultanés de seuils élevés, appelés, dans le cas de la dimension infinie, des Processus de Pareto Généralisés (GPP, *Generalized Pareto Process*) (Ferreira et de Haan, 2012). Ces derniers sont la contrepartie des processus max-stables, obtenue en considérant les dépassements de seuils élevés à la place des maxima de blocs de grande taille. Les processus max-stables ont reçu une attention considérable ces dernières années (De Haan, 1984; Smith, 1990; Schlather, 2002; Westra et Sisson, 2011; Padoan *et al.*, 2010). Les trajectoires de tels processus (les réalisations), peuvent légitimement servir de modèle pour les maxima de blocs de grande taille, calculés « composante par composante ». Ainsi, elles ne correspondent pas à des réalisations observées du processus sous-jacent. Cette approche est adaptée à certaines questions ayant trait à la gestion des risques naturels, l'exemple typique étant de connaître la probabilité de dépassement par le niveau de la mer, d'une digue côtière, en au moins un endroit de la digue, au moins une fois sur un nombre d'années n grand. En revanche, pour certaines applications, la question est de modéliser (estimer et simuler) les réalisations *effectives* de champs spatiaux prenant de très grandes valeurs.

Revenons à l'exemple classique des pluies extrêmes à l'échelle d'une région. L'occurrence simultanée de précipitations intenses en plusieurs points d'intérêt est susceptible de provoquer les dommages les plus importants. Par ailleurs, pour les besoins de la modélisation météorologique ou hydrologique, le recours aux simulateurs de pluies est fréquent. Pour simuler des pluies éventuellement intenses, la connaissance de la structure spatiale des réalisations effectives des champs de pluies, conditionnellement à un excès de seuil (à définir), est nécessaire.

La simulation directe de tels champs est une question délicate, et une solution partielle consiste à générer un processus max-stable appartenant au même domaine d'attraction (Buishand *et al.*, 2008). Dans le cadre des processus continus, la difficulté de la simulation directe à partir de la mesure angulaire vient de l'obligation pour l'« angle » d'appartenir à la sphère unité. En d'autres termes, l'emploi de la norme sup pour définir la sphère unité dans l'espace des processus oblige à simuler un processus $V(s)_{s\in S}$ tel que $\sup_{s\in S} V(s) = 1$ presque sûrement. Cette contrainte est liée au choix du seuil, ce qui revient, dans le cas spatial, à définir une fonction de coût $\mathcal{L}$ sur l'espace des trajectoires et à déclarer « extrêmes » les réalisations correspondant à de grandes valeurs de $\mathcal{L}$. Une première étape vers un affaiblissement de la

contrainte précédente consiste à admettre plusieurs fonctions de coût, sous une hypothèse de continuité par rapport à la norme uniforme. Dombry et Ribatet (2013) montrent que dans ce cas, les mesures spectrales (et la loi des angles) correspondant aux différents coûts se déduisent facilement les unes des autres. Une question presque ouverte consiste à se demander s'il est possible d'abandonner l'hypothèse de continuité des trajectoires et de travailler à la place sur un espace de mesures.

Cette piste est partiellement explorée par Boldi (2004) dans sa thèse de doctorat. Après avoir divisé la région d'intérêt S en une partition formée de sous-régions homogènes $S_1, \ldots, S_d$, un mélange de Dirichlet en dimension d est ajusté. Le principal problème est qu'une fois la partition imposée, l'estimation concerne seulement les paramètres $\alpha_m(S_1), \ldots, \alpha_m(S_d)$ ($m \leq k$, pour un mélange de processus de Dirichlet à k composants $\alpha_1, \ldots, \alpha_k$), de sorte qu'on ne peut faire de prédiction sur une région A n'appartenant pas à la partition d'origine. Une possibilité serait de spécifier un modèle sur les α_m, par exemple en imposant une forme gaussienne ou triangulaire, puis d'inférer les hyper-paramètres gouvernant l'échelle et la localisation, de sorte que $\alpha_m(A)$ serait accessible pour toute région A. L'inconvénient apparent est que cela restreint l'espace des mesures angulaires possibles, mais la possibilité d'augmenter le nombre de composants du mélange est prometteuse pour contourner cette difficulté.

D'un point de vue théorique, le cadre des extrêmes multivariés ne s'étend pas tel quel aux espaces de mesures, notamment parce que l'espace des mesures régulières sur un ouvert de $\mathbf{R}^2$ n'est pas localement compact, propriété pourtant nécessaire dans l'établissement classique de l'équivalence entre variation régulière, convergence des excès vers un processus de Poisson et attraction des maxima. Cependant la notion de variation régulières des mesures sur $\mathbf{R}^d$ s'étend à des mesures définies sur tout espace métrique séparable, complet (Hult et Lindskog, 2006), en particulier l'espace des mesures positives régulières $M_+(\boldsymbol{E})$. Les travaux de Hult et Lindskog (2006); Ferreira et de Haan (2012); Dombry et Ribatet (2013) se concentrent sur les espaces de processus continus. La direction des mesures aléatoires reste encore à explorer.

Annexe A

Note sur la formule de Bayes

L'objectif de cette note est de montrer que l'hypothèse 1 faite en section 1.3.1 est vérifiée sous une condition très faible. On verra que la mesurabilité de l'application partielle $p_x : \theta \mapsto p_\theta(x)$ a lieu presque partout, dans un sens à préciser, sous une condition très générale sur l'espace des observations, à savoir

Hypothèse 2.
La tribu des observations $\mathcal{X}$ est engendré par une collection dénombrable d'ensembles.

Cette dernière hypothèse est en particulier vérifiée si l'espace des observations est CMS. Notons que l'hypothèse de régularité assurant l'existence de la loi conditionnelle concerne $\mathcal{X}$ et non $\mathcal{T}$, comme on pourrait s'y attendre au vu du théorème classique de désintégration (voir *e.g.* Kallenberg, 1997, théorème 5.3).

On va montrer que $p_\theta(x)$ est presque partout égale à une dérivée de Radon-Nikodym d'une mesure jointe sur $\boldsymbol{X} \times \boldsymbol{\Theta}$ par rapport à une mesure de référence, comme suggéré dans Schervish (1995), problème 1.9. Dans la suite,les σ-algèbres sont toujours considérées comme complètes pour la mesure considérée, de sorte que les sous-ensembles d'ensembles négligeables sont toujours mesurables.

Soit λ la mesure sur $\mathcal{X} \otimes \mathcal{T}$ définie par

$$\forall A \in \mathcal{X},\ \forall B \in \mathcal{T}, \quad \lambda(A \times B) = \int_B \mathrm{P}_\theta(A)\,\mathrm{d}\pi(\theta)\,.$$

Lemme 1. *La mesure jointe λ est absolument continue par rapport à $\nu \times \pi$.*

Démonstration. Soit $G \in \mathcal{X} \otimes \mathcal{T}$ et, pour $\theta \in \boldsymbol{\Theta}$, soit G_θ la θ-section $\{x \in$

$\boldsymbol{X} : (x, \theta) \in G\}$. On a immédiatement

$$\begin{aligned}
\nu \times \pi(G) = 0 \quad &\Rightarrow \int_{\boldsymbol{\Theta}} \int_{G_\theta} \mathrm{d}\nu(x)\, \mathrm{d}\pi(\theta) = 0 \quad \text{(Tonelli-Fubini)} \\
&\Rightarrow \int_{\boldsymbol{\Theta}} \nu(G_\theta)\, \mathrm{d}\pi(\theta) = 0 \\
&\Rightarrow \nu(G_\theta) = 0\,, \quad \pi\text{-p.s.} \\
&\Rightarrow \mathrm{P}_\theta(G_\theta) = 0\,, \quad \pi\text{-p.s.} \quad (\text{since } \mathrm{P}_\theta \ll \nu\,, \pi\text{-p.s. }) \\
&\Rightarrow \int_{\boldsymbol{\Theta}} \mathrm{P}_\theta(G_\theta)\, \mathrm{d}\pi(\theta) = 0 \\
&\Rightarrow \lambda(G) = 0\,.
\end{aligned}$$

□

La dérivée de Radon de λ par rapport à $\nu \times \pi$ est donc bien définie. Notons-la $f(x, \theta)$. Par définition, f est $\mathcal{X} \otimes \mathcal{T}$-mesurable et $f_x : \theta \mapsto f(x, \theta)$ est ν-presque partout $\mathcal{T}$-mesurable (Fubini-Tonelli pour mesures complètes, voir par exemple Folland, 1984, théorème 2.39). Le résultat principal peut maintenant être énoncé.

Théorème 9. *La dérivée de Radon $f(x, \theta) = \frac{\mathrm{d}\lambda}{\mathrm{d}\nu \times \pi}(x, \theta)$ et la vraisemblance $p_\theta(x)$ coïncident, dans le sens suivant : Il existe un ensemble π-négligeable $N \subset \boldsymbol{\Theta}$, tel que, pour tout $\theta \in \boldsymbol{\Theta} \setminus N$, pour tout x en dehors d'un ensemble ν-négligeable N_θ, $f(x, \theta) = p_\theta(x)$.*

Ainsi, le modèle $\mathcal{P}_{\boldsymbol{\Theta}}$ est π-presque-partout inchangé si l'on remplace la vraisemblance p_θ par sa version $\mathcal{T}$-mesurable f_θ.

La preuve utilise le lemme suivant

Lemme 2. $\forall A \in \mathcal{X}, \exists N_A \in \mathcal{T},\ \pi(N_A) = 0$, *tel que*

$$\forall \theta \in N_A^c, \int_A p_\theta(x)\, \mathrm{d}\nu(x) = \int_A f(x, \theta)\, \mathrm{d}\nu(x)$$

Démonstration. Soit A comme dans l'énoncé et $B \in \mathcal{T}$.
D'une part,

$$\begin{aligned}
\lambda(A \times B) &= \int_B \int_A \mathrm{dP}_\theta(x)\, \mathrm{d}\pi(\theta) \\
&= \int_B \int_A p_\theta(x)\, \mathrm{d}\nu(x)\, \mathrm{d}\pi(\theta) \quad (P_\theta \ll \nu, \quad \pi\text{-p.s.}) \qquad \text{(A.1)}
\end{aligned}$$

D'autre part,

$$\begin{aligned}
\lambda(A \times B) &= \int_{\boldsymbol{\Theta} \times \boldsymbol{X}} \mathbb{1}_{A \times B}(x, \theta) \frac{\mathrm{d}\lambda}{\mathrm{d}\nu \times \pi}(x, \theta)\ \mathrm{d}(\nu \times \pi)(x, \theta) \\
&= \int_B \int_A f(x, \theta)\, \mathrm{d}\nu(x)\, \mathrm{d}\pi(\theta) \quad \text{(Tonelli-Fubini)} \qquad \text{(A.2)}
\end{aligned}$$

En combinant (A.2) et (A.1), on obtient

$$\int_B \left[\int_A p_\theta(x)\,\mathrm{d}\nu(x)\right]\,\mathrm{d}\pi(\theta) = \int_B \left[\int_A f(x,\theta)\,\mathrm{d}\nu(x)\right]\,\mathrm{d}\pi(\theta)$$

Ceci étant vrai pour tout $B \in \mathcal{T}$, le résultat suit. □

La preuve du théorème 9 utilise le lemme précédent et un argument de classe monotone (ou théorème $\lambda - \pi$ de Dynkin, voir par exemple Kallenberg, 1997, théorème 1.1)

Preuve du Théorème 9. Il suffit de montrer que l'ensemble de mesure nulle N_A du lemme 2 ne dépend pas de A. D'après l'hypothèse 2, il existe une collection dénombrable $\mathcal{A} = \{A_n, n \in \mathbf{N}\}$ de parties de $\boldsymbol{X}$ qui engendre $\mathcal{X}$. Soit $\mathcal{C}$ l'ensemble des intersections finies d'éléments de $\mathcal{A}$. La famille $\mathcal{C}$ est encore dénombrable, et l'on peut écrire $\mathcal{C} = \{C_n, n \in \mathbf{N}\}$. De plus, $\mathcal{C}$ est un π-système (elle est stable par intersections finies) qui engendre $\mathcal{X}$. D'après le lemme 2, pour tout $n \in \mathbf{N}$, il existe $N_n \in \mathcal{T}$, $\pi(N_n) = 0$, tel que

$$\forall \theta \notin N_n, \int_{C_n} f(x,\theta)\,\mathrm{d}\nu(x)\,\mathrm{d}\pi(\theta) = \int_{C_n} p_\theta(x)\,\mathrm{d}\nu(x)\,\mathrm{d}\pi(\theta)\,.$$

Posons $\bar{N} = \bigcup_{n\in\mathbf{N}} C_n$. À nouveau, $\pi(\bar{N}) = 0$.

Soit $\mathcal{D}$ l'ensemble des parties $D \subset \mathcal{X}$ ayant la propriété désirée, à savoir tels que

$$\forall \theta \notin \bar{N}, \quad \int_D f(x,\theta)\,\mathrm{d}\nu(x)\,\mathrm{d}\pi(\theta) = \int_D p_\theta(x)\,\mathrm{d}\nu(x)\,\mathrm{d}\pi(\theta)\,.$$

Par construction, on a $\mathcal{C} \subset \mathcal{D}$. De plus, on vérifie facilement que $\mathcal{D}$ est stable par unions dénombrables et différences propres, et que $\mathcal{X} \in \mathcal{D}$. Ainsi, $\mathcal{D}$ est une classe monotone, et par le lemme de classe monotone, $\mathcal{D}$ contient la tribu engendrée $\sigma(\mathcal{C}) = \mathcal{T}$.

En conclusion, pour tout $\theta \notin \bar{N}$, pour tout $A \in \mathcal{X}$,

$$\int_A f(x,\theta)\,\mathrm{d}\nu(x)\,\mathrm{d}\pi(\theta) = \int_A p_\theta(x)\,\mathrm{d}\nu(x)\,\mathrm{d}\pi(\theta)\,,$$

de sorte que, pour tout $\theta \notin \bar{N}$, les mesures sur $\mathcal{X}$ de densités respectives $p_\theta(\,\cdot\,)$ et $f(\,\cdot\,,\theta)$ par rapport à ν définissent la même mesure, d'où le résultat. □

Liste des symboles

$(\cdot)_+$	Partie positive, page 10
$\bigvee$	Opérateur maximum (terme à terme), page 11
$\mathcal{C}_c^+$	Ensemble des fonctions continues, positives, à support compact, page 18
$\mathscr{C}(F)$	Ensemble des points de continuité de la fonction F, page 12
DA(G)	Domaine d'attraction (des maxima) de la loi G, page 14
DAE(Q)	Domaine d'attraction des excès de la loi Q, page 12
ℓ	Mesure de Lebesgue sur $\mathbf{R}$ ou $\mathbf{R}^d$, page 11
$\mathcal{C}_b$	Ensemble des fonctions continues bornées, page 11
$\mathbf{N}$	Ensemble des entiers naturels, page 10
$\mathbf{R}$	Droite réelle, page 10
$\xrightarrow{d}$	Convergence faible des mesures de probabilité, page 11
$\xrightarrow{v}$	Convergence vague des mesures de Radon, page 18
$f^{\leftarrow}$	Inverse généralisée de la fonction croissante f, page 16
M_+	Ensemble des mesures boréliennes, positives, finies sur les compacts (mesures de Radon), page 18
DM	Dirichlet mixture, page 24
f.d.r.	Fonction de répartition, page 11
CMS	Espace métrique séparable et complet, page 18
GEV	Generalized Extreme Value distribution, page 14
GPD	Generalized Pareto Distribution, page 13
POT	Peaks-over-threshold, caractérisation des extrêmes par la loi des excès au-dessus de seuils élevés, page 5

Bibliographie

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. bn petrov and f. czàke, eds. *In Second International Symposium on Information Theory. Akademiai Kiadó, Budapest.*

Balkema, A. A. et De Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, pages 792–804.

Beirlant, J., Goegebeur, Y., Segers, J. et Teugels, J. (2004). *Statistics of extremes : Theory and applications.* John Wiley & Sons : New York.

Berger, J. O. et Wolpert, R. L. (1988). The likelihood principle.

Berk, R. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58.

Blanchet, J. et Davison, A. C. (2011). Spatial modeling of extreme snow depth. *The Annals of Applied Statistics*, 5(3):1699–1725.

Boldi, M. (2004). *Mixture models for multivariate extremes.* Thèse de doctorat, Ecole Polytechnique Federale de Lausanne.

Boldi, M.-O. et Davison, A. C. (2007). A mixture model for multivariate extremes. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 69(2):217–229.

Buishand, T., de Haan, L. et Zhou, C. (2008). On spatial extremes : with application to a rainfall problem. *The Annals of Applied Statistics*, pages 624–642.

Chang, J. T. et Pollard, D. (1997). Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317.

Coles, S. (2001). *An introduction to statistical modeling of extreme values.* Springer Verlag.

Coles, S. et Tawn, J. (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 377–392.

Coles, S. G. et Powell, E. A. (1996). Bayesian methods in extreme value modelling : a review and new developments. *International Statistical Review/Revue Internationale de Statistique*, pages 119–136.

Cooley, D., Nychka, D. et Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479): 824–840.

Daley, D. J. et Vere-Jones, D. (2007). *An introduction to the theory of point processes : volume II : general theory and structure*, volume 2. Springer.

Davison, A. et Smith, R. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 393–442.

de Carvalho, M., Oumow, B., Segers, J. et Warchoł, M. (2013). A euclidean likelihood estimator for bivariate tail dependence. *Communications in Statistics-Theory and Methods*, 42(7).

de Haan, L. (1970). On regular variation and its application to the weak convergence of sample extremes. *MC Tracts*, 32:1–124.

de Haan, L. (1976). Sample extremes : an elementary introduction. *Statistica Neerlandica*, 30(4):161–172.

De Haan, L. (1984). A spectral representation for max-stable processes. *The annals of probability*, pages 1194–1204.

de Haan, L. et Ferreira, A. (2006). *Extreme Value Theory, An Introduction.* Springer Series in Operations Research and Financial Engineering.

Dombry, C. et Ribatet, M. (2013). Functional regular variations, pareto processes and peaks over threshold.

Einmahl, J., de Haan, L. et Piterbarg, V. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *The Annals of Statistics*, 29(5):1401–1423.

Einmahl, J. et Segers, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics*, 37(5B):2953–2989.

Embrechts, P., Klüppelberg, C. et Mikosch, T. (1997). *Modelling extremal events : for insurance and finance*, volume 33. Springer Verlag.

Ferreira, A. et de Haan, L. (2012). The generalized pareto process ; with a view towards application and simulation. *arXiv preprint arXiv :1203.2551v2.*

Fisher, R. A. et Tippett, L. (1928). Limiting forms of the frequency distribution of the largest of smallest member of a sample. *In Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190.

Folland, G. B. (1984). *Real analysis : modern techniques and their applications*, volume 2. Wiley New York.

Ghosh, J. et Ramamoorthi, R. (2003). *Bayesian nonparametrics.* Springer.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *The Annals of Mathematics*, 44(3):423–453.

Gneiting, T. et Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Guillotte, S., Perron, F. et Segers, J. (2011). Non-parametric Bayesian inference on bivariate extremes. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*.

Gumbel, E. J. (1958). *Statistics of extremes*. Columbia University Press (New York).

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hoeting, J., Madigan, D., Raftery, A. et Volinsky, C. (1999). Bayesian model averaging : A tutorial. *Statistical science*, 14(4):382–401.

Hult, H. et Lindskog, F. (2006). Regular variation for measures on metric spaces. *Publ. Inst. Math.(Beograd)(NS)*, 80(94):121–140.

Hüsler, J. et Li, D. (2009). Testing asymptotic independence in bivariate extremes. *Journal of Statistical Planning and Inference*, 139(3):990–998.

Kallenberg, O. (1997). *Foundations of modern probability*. Springer Verlag.

Leadbetter, M. R., Lindgren, G. et Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer-Verlag, New York.

Lorentz, G. G. (1953). *Bernstein polynomials*. American Mathematical Soc.

Madigan, D. et Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

Meyn, S., Tweedie, R. et Glynn, P. (1993). *Markov chains and stochastic stability*. Springer London et al.

Neppel, L., Renard, B., Lang, M., Ayral, P., Coeur, D., Gaume, E., Jacob, N., Payrastre, O., Pobanz, K. et Vinet, F. (2010). Flood frequency analysis using historical data : accounting for random and systematic errors. *Hydrological Sciences Journal–Journal des Sciences Hydrologiques*, 55(2):192–208.

Padoan, S. A., Ribatet, M. et Sisson, S. A. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105(489).

Parent, E. et Bernier, J. (2003). Bayesian pot modeling for historical data. *Journal of hydrology*, 274(1):95–108.

Petrone, S. et Wasserman, L. (2002). Consistency of bernstein polynomial posteriors. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(1):79–100.

Pickands, J. I. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131.

Raftery, A., Gneiting, T., Balabdaoui, F. et Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174.

Resnick, S. (1987). *Extreme values, regular variation, and point processes, volume 4 of Applied Probability. A Series of the Applied Probability Trust*. Springer-Verlag, New York.

Ribatet, M., Sauquet, E., Grésillon, J.-M. et Ouarda, T. B. (2007). A regional bayesian pot model for flood frequency analysis. *Stochastic Environmental Research and Risk Assessment*, 21(4):327–339.

Robert, C. (2007). *The Bayesian choice : from decision-theoretic foundations to computational implementation*. Springer Verlag.

Robert, C. et Casella, G. (2004). *Monte Carlo statistical methods*. Springer Verlag.

Robert, C. et Casella, G. (2010). *Introducing Monte Carlo Methods with R*. Springer Verlag.

Roberts, G. et Rosenthal, J. (2004). General state space Markov chains and mcmc algorithms. *Probability Surveys*, 1:20–71.

Roberts, G. et Rosenthal, J. (2006). Harris recurrence of metropolis-within-gibbs and trans-dimensional Markov chains. *The Annals of Applied Probability*, 16(4):2123–2139.

Roberts, G. et Smith, A. (1994). Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. *Stochastic processes and their applications*, 49(2):207–216.

Sabourin, A. et Naveau, P. (2013). Bayesian dirichlet mixture model for multivariate extremes : A re-parametrization. *Computational Statistics & Data Analysis*.

Sabourin, A., Naveau, P. et Fougères, A.-L. (2013). Bayesian model averaging for multivariate extremes. *Extremes*, 16(3):325–350.

Schervish, M. J. (1995). *Theory of statistics*. Springer Series in Statistics.

Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes*, 5(1): 33–44.

Schwartz, L. (1965). On Bayes procedures. *Probability Theory and Related Fields*, 4:10–26. 10.1007/BF00535479.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Smith, R. L. (1990). Max-stable processes and spatial extremes. *Unpublished manuscript, Univer*.

Tsai, Y.-L., Dupuis, D. J. et Murdoch, D. J. (2013). A robust test for asymptotic independence of bivariate extremes. *Statistics*, 47(1):172–183.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Westra, S. et Sisson, S. A. (2011). Detection of non-stationarity in precipitation extremes using a max-stable process model. *Journal of Hydrology*, 406(1):119–128.