HAL
open science

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

Anne Sabourin, Philippe Naveau

To cite this version:

Anne Sabourin, Philippe Naveau. Bayesian Dirichlet mixture model for multivariate extremes: a re-parametrization. Computational Statistics and Data Analysis, 2013, 10.1016/j.csda.2013.04.021 . hal-00880879

HAL Id: hal-00880879

https://hal.science/hal-00880879v1

Submitted on 11 Nov 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

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

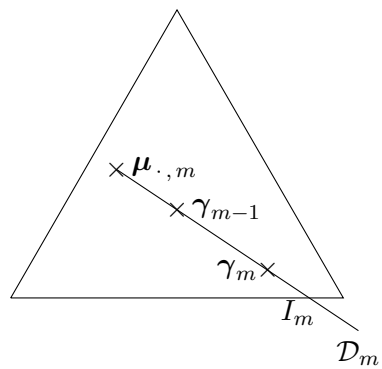

Figure 1: Sequential construction of the partial centers of mass on the two-dimensional simplex $\mathbf{S}_3$ at step m. The simplex points γ_m, γ_{m-1} and m^{th} mean vector $\boldsymbol{\mu}_{\cdot,m}$, as defined in Proposition 1, belong to a common line $\mathcal{D}_m$ and γ_{m-1} lies between 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 $p_{1:k}$.

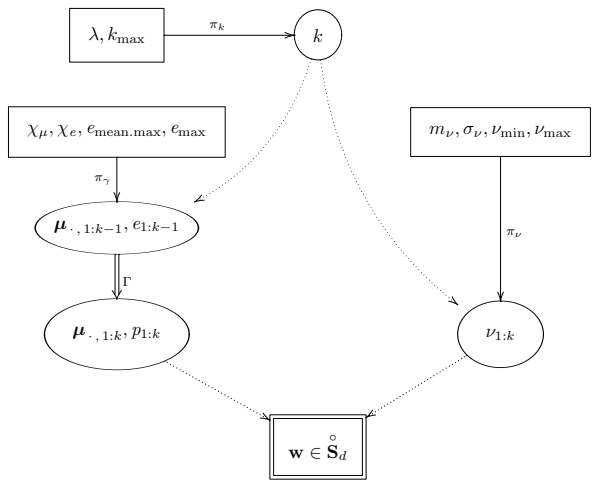

Figure 2: Representation of the conditional dependencies of the DM Bayesian model as a Directed acyclic graph. Hyper-parameters appear in simple square frames, parameters in oval frames and observations in a double square frame. Simple arrows denote probabilistic relations whereas double arrows stand 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: $\hat{\delta}^{\mathrm{nonP}}(g) = \frac{1}{\sqrt{n}}\left[\widehat{(g^2)}^{\mathrm{nonP}} - \left(\hat{g}^{\mathrm{nonP}}\right)^2\right]^{1/2}$.

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

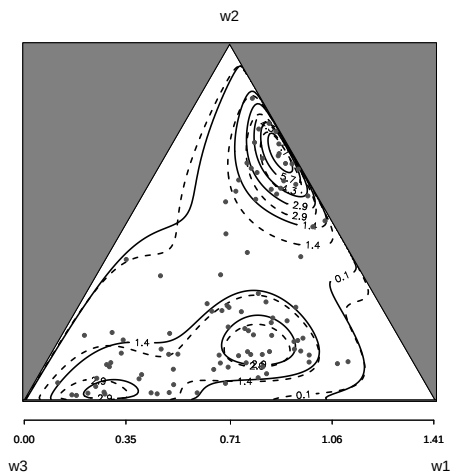

Figure 3: Predictive angular density contours (solid lines) obtained *via* the M-DMalgorithm, on the two-dimensional simplex $\mathbf{S}_3$, inferred with 100 simulated points (Gray points) simulated from the true density (dotted lines) defined by (20).

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

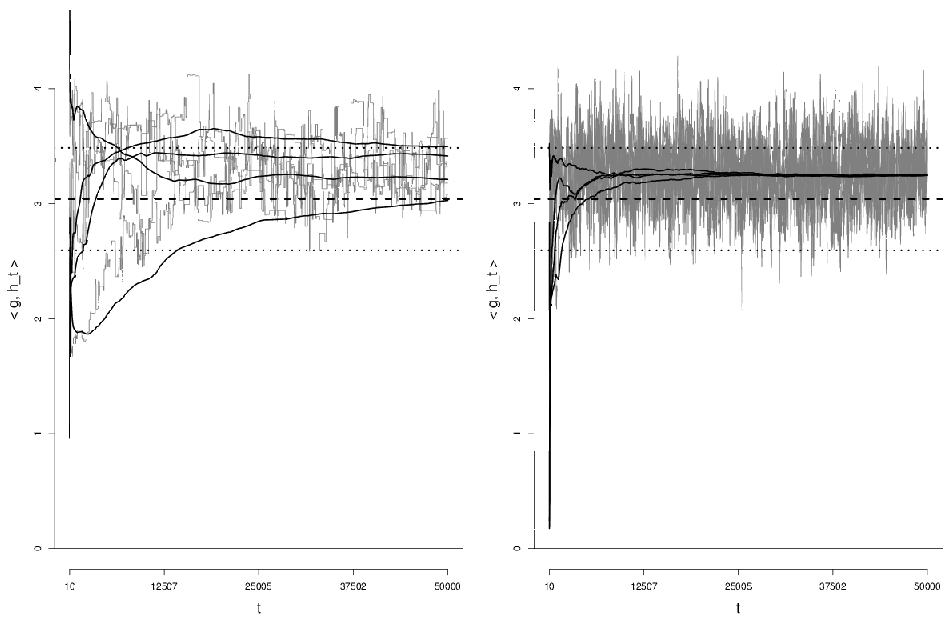

Figure 4: Convergence monitoring with three-dimensional data, with four parallel chains in each model. Integration of the densities generated by the original DM model (left panel) and by the re-parametrized version (right panel) against a Dirichlet density with parameter $\nu\ \boldsymbol{\mu} \simeq (7.67, 2.72, 4.60)$.
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 $(0, 1)$.

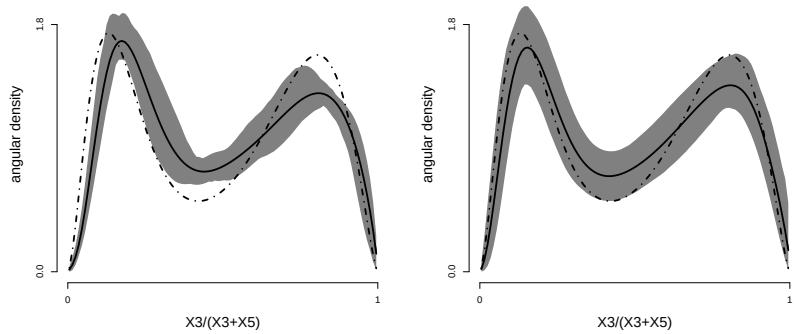

Figure 5: Simulated five-dimensional data (100 points): Bi-variate angular posterior predictive densities for the pair $(3, 5)$. Left panel: Original algorithm; Right panel: re-parametrized version. Dash-dotted line: true density; solid line: posterior predictive; Gray area: posterior credible set at 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.

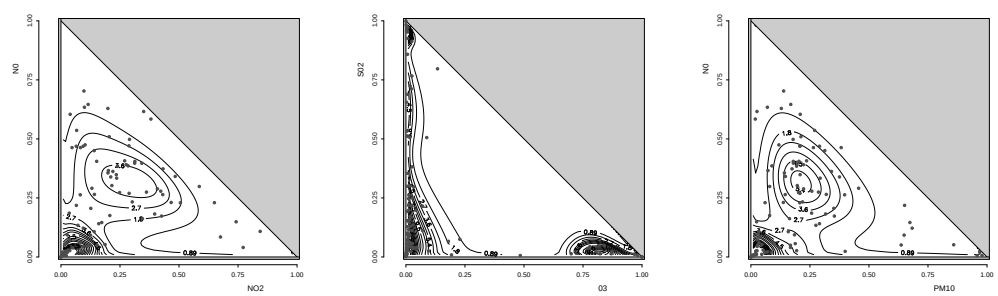

Figure 6: Five dimensional Leeds data set: posterior predictive density. Black lines: projections of the predictive angular density defined on the four-dimensional simplex $\mathbf{S}_5$ onto the two-dimensional faces. Gray dots: projections of the 100 points with greatest L^1 norm.

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.

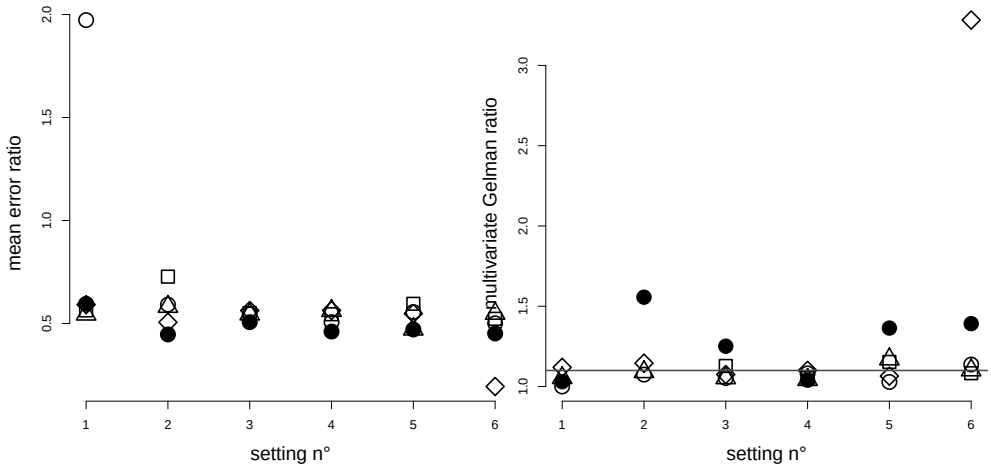

Figure 7: Influence of the prior on the quality of the fit (left panel, mean error ratio $\bar{r}^{\mathrm{DM}}$), and on the chains' mixing properties (right panel, multivariate potential scale reduction factor R_G). $\bullet$, simplified prior on $(\boldsymbol{\mu}, e)$, influence of σ_ν (variance 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) ; $\square$, dependent prior on $(\boldsymbol{\mu}, e)$, influence of σ_ν, same values for σ_ν; $\circ$, influence of λ (mean number of mixture components), $\lambda \in \{1, 3, 5, 7, 10, 12\}$; $\diamond$, influence of χ_μ (concentration of mean vectors), $\chi_\mu \in \{0.5, 1, 1.5, 2, 4, 8\}$; $\triangle$, 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-

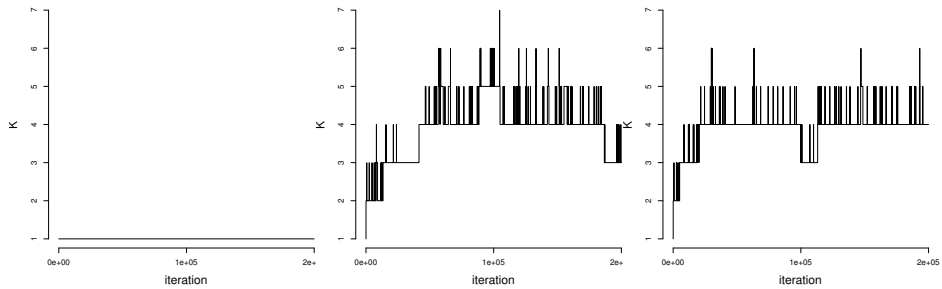

Figure 8: Influence of hyper-parameter λ on the number k of mixture components generated by the reversible-jump sampler. From left to right, evolution of k in the MCMC run, for $\lambda \in \{1, 3, 12\}$, and a 100-points data set issues from four-components Dirichlet mixture.

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

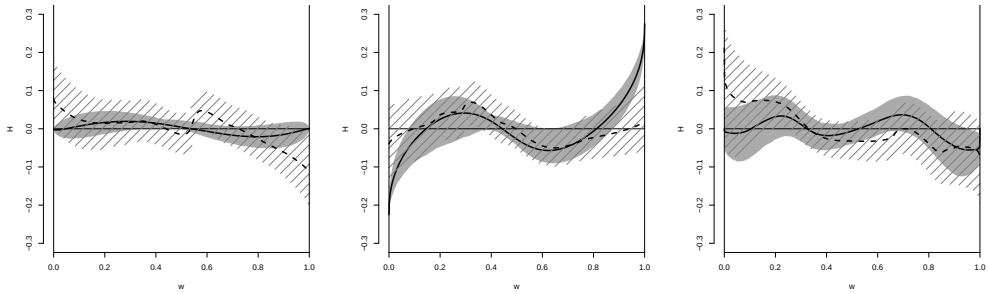

Figure 9: Error of the predictive angular *cdf* (solid lines) on the segment [0,1]. From left to right: data from a Logistic, an Asymmetric logistic and from a DM distribution. Solid line and gray area: Dirichlet Mixture mean estimate and $0.1 - 0.9$ posterior quantiles; dashed line and dashed area: *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, $\mathcal{C}_m = \left\{i \in \{1,\ldots,d\} \ :\ \gamma_{i,m-1} - \mu_{i,m} < 0\right\}$. 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

$$\begin{aligned}
T_m &= \sup\left\{t \geq 0 \ :\quad \forall i \in \mathcal{C}_m,\ \gamma_{i,m} + t\,(\gamma_{i,m} - \mu_{i,m}) > 0\right\} \\
&= \sup\left\{t \geq 0 \ :\quad t < \min_{i\in\mathcal{C}_m}\left(\frac{\gamma_{i,m}}{\mu_{i,m} - \gamma_{i,m}}\right)\right\} \\
&= \min_{i\in\mathcal{C}_m}\left(\frac{\gamma_{i,m}}{\mu_{i,m} - \gamma_{i,m}}\right).
\end{aligned}$$

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 $\theta = \left(\boldsymbol{\mu}_{\cdot,1:k-1}, e_{1:k-1}, \nu_{1:k}\right) \in \boldsymbol{\Theta}_k$. 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

$$
\begin{aligned}
a_{\min} &= \min\left\{\mu_{i,m}\,\nu_m : \; m \le k, i \le d, (\,\boldsymbol{\mu}, \mathbf{e}, \boldsymbol{\nu}) \in \bar{B}_\epsilon\right\} > 0\,, \\
a_{\max} &= \max\left\{\mu_{i,m}\,\nu_m : \; m \le k, i \le d, (\,\boldsymbol{\mu}, \mathbf{e}, \boldsymbol{\nu}) \in \bar{B}_\epsilon\right\}, \\
D_{\min} &= \min\left\{\frac{\Gamma(\nu_m)}{\prod_{i=1}^d \Gamma(\mu_{i,m}\,\nu_m)} : m \le k, i \le d, (\,\boldsymbol{\mu}, \mathbf{e}, \boldsymbol{\nu}) \in \bar{B}_\epsilon\right\} > 0\,, \\
D_{\max} &= \max\left\{\frac{\Gamma(\nu_m)}{\prod_{i=1}^d \Gamma(\mu_{i,m}\,\nu_m)} : m \le k, i \le d, (\,\boldsymbol{\mu}, \mathbf{e}, \boldsymbol{\nu}) \in \bar{B}_\epsilon\right\} \; .
\end{aligned}
$$

Note that, by compactness of $\bar{B}_\epsilon$, the *extrema* are reached, which ensures positivity of the *infima*. Hence, $\forall(\,\boldsymbol{\mu}, \mathbf{e}, \boldsymbol{\nu}) \in \bar{B}_\epsilon, \forall \mathbf{w} \in \overset{\circ}{\mathbf{S}}_d$, $\forall m \le k$,

$$
0 < D_{\min} \prod_{1 \le i \le d} w_i^{a_{\max}-1} \le \mathrm{diri}(\mathbf{w} \mid \boldsymbol{\mu}_{\cdot,m}, \nu_m) \le D_{\max} \prod_{1 \le i \le d} w_i^{a_{\min}-1} \; .
$$

By convex combination, we also have, $\forall \theta \in \bar{B}_\epsilon, \forall \mathbf{w} \in \overset{\circ}{\mathbf{S}}_d$,

$$
0 < D_{\min} \prod_i w_i^{a_{\max}-1} \le h_\theta(\mathbf{w}) \le D_{\max} \prod_i w_i^{a_{\min}-1} \; .
$$

Whence, by monotonicity of the log function, $\exists D_1, D_2 > 0$,

$$
D_1 + (a_{\max} - 1) \sum_i \log(w_i) \le \log\left(h_\theta(\mathbf{w})\right) \le D_2 + (a_{\min} - 1) \sum_i \log(w_i) \; .
$$

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 $\mathrm{Im}(K_m, \theta) \subset \mathbf{\Theta}$ the image of $K_m(\theta, \cdot)$, η_m the reference Lebesgue measure on $\mathrm{Im}(K_m, \theta)$, 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.