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

A. SABOURIN[a,1], 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 among large events is characterized by an angular measure on the positive quadrant 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 rendering Bayesian inference very challenging in dimension greater than three. In this paper, a new parametrization is proposed which is unconstrained and allows for a natural prior specification, which posterior consistency is verified. A reversible-jump algorithm is implemented to approximate the posterior and tested up to dimension five. In this non identifiable setting, convergence assessment 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 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 is a strong argument in favor of 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 is proven, but the intricate form of the limit law does not provide, to our understanding, a simple way to derive asymptotic confidence bounds. Within a Bayesian framework, the only article (to your knowledge) dealing with a fully non-parametric model was restricted to the bi-variate case (Guillotte et al., 2011). In a semi-parametric context, Boldi and Davison (2007) proposed a Dirichlet mixture (DM) model with varying number of mixture components, which 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) 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

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

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 consistent prior in a relatively simple way (see Section 3). A trans-dimensional *Metropolis-within-Gibbs* algorithm is implemented (see 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 set (see Section 7).

Theoretical ergodicity properties of the algorithm are established in Section 5. In Section 6, the important issue of empirical convergence assessment is investigated. As it is the case for any other mixture model, the parameters of the mixture are not identifiable, and the monitored quantity cannot be the parameter itself. 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 Section 7, a simulation study is performed with two- and five- dimensional data sets, in order to compare our algorithm with Boldi and Davison's one, in terms of mixing properties and precision accuracy. We also fit our model to air quality measurements[1] recorded in the city of Leeds, UK, during the winter season, years 1994-1998. This data set was already studied by Cooley et al. (2010), Heffernan and Tawn (2004), Boldi and Davison (2007) and Sabourin et al. (In press). We comment our results with respect to Boldi and Davison (2007)'s approach.

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 $\mathbf{S}_d = \{\mathbf{w} = (w_1, \cdots, w_d) \ : w_i \geq 0 \ \ w_1 + \cdots + w_d = 1\}$.

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.* there exists a non degenerate *cdf* G such 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

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

[1] Available at `http://www.airquality.co.uk`

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 $\mathbf{p} = (p_1, \ldots, p_k)$ summing to one, concentration vector $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)$ and mean matrix $\boldsymbol{\mu} = (\boldsymbol{\mu}_{\cdot,1}, \ldots, \boldsymbol{\mu}_{\cdot,k})$ where $\boldsymbol{\mu}_{\cdot,m} = (\mu_{1,m}, \ldots, \mu_{d,m})$ is the mean vector for the m^{th} 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[2] proposed and studied by Boldi and Davison (2007) as a disjoint union:

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

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 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. 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.

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

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 $m \in \{0, \ldots, k-1\}$, let $\boldsymbol{\gamma}_m$ be the center of mass of the $k-m+1$ last components

$$\boldsymbol{\gamma}_m = \rho_m^{-1} \sum_{j=m+1}^{k} p_j \boldsymbol{\mu}_{\cdot,j}, \tag{4}$$

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 $\boldsymbol{\gamma}_0 = (1/d, \ldots, 1/d)$. By associativity of the center of mass, 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 $m = 0$, 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

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

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 $\boldsymbol{\gamma}_1$.

The argument can be repeated to obtain subsequent centers of mass $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{k-1}$ and weights $p_1, \ldots, p_{k-1}, \rho_1, \ldots, \rho_{k-1}$, given $k-1$ Dirichlet mean vectors $\boldsymbol{\mu}_{\cdot,1:k-1}$ and eccentricities $e_{1:k-1}$, *via*

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

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 e_m small, $\boldsymbol{\gamma}_{m-1}$ 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

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

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.1 that

$$T_m = \min_{i \in C_m} \frac{\boldsymbol{\gamma}_{i,m-1}}{\mu_{i,m} - \boldsymbol{\gamma}_{i,m-1}}, \tag{8}$$

where C_m is the index set $\{i \in \{1, \ldots, d\} \,:\, \gamma_{i,m-1} - \mu_{i,m} < 0\}$.

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 *$(e_1, \ldots, e_{k-1}) \in (0,1)^{k-1}$ 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 $1 \leq m \leq k-1$,*

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

Defining the last mean vector $\boldsymbol{\mu}_{\cdot,k}$ and weight p_k by $\boldsymbol{\mu}_{\cdot,k} = \boldsymbol{\gamma}_{k-1}$ 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.*

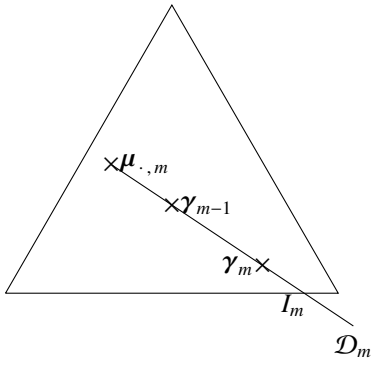

Figure 1: Sequential construction of the partial centers of mass on the two-dimensional simplex $\mathbf{S}_3$ at step m. The simplex points γ_m, γ_{m-1} and m^{th} mean vector $\boldsymbol{\mu}_{\cdot,m}$, as defined in Proposition 1, belong to a common line $\mathcal{D}_m$ and γ_{m-1} lies between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

The *unconstrained parameter space* for the DM model can now be defined as a disjoint union

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

For $k \geq 1$, we introduce the re-parametrization maps for k-mixtures $\Gamma_k : \theta \in \boldsymbol{\Theta}_k \mapsto (\boldsymbol{\mu}_{\cdot,1:k}, p_{1:k}, \nu_{1:k}) \in \Psi_k$, which allows to define

$$\Gamma : \boldsymbol{\Theta} \longrightarrow \Psi$$
$$\theta \in \boldsymbol{\Theta}_k \longmapsto \Gamma_k(\theta) \in \Psi_k$$

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

$$\mathbf{\Theta}_B = \prod_{k=1}^{k_{\max}} \mathbf{S}_d^{k-1} \times [0, e_{\max}]^{k-1} \times [\nu_{\min}, \nu_{\max}]^k \tag{9}$$

with, typically, $k_{\max} = 15$, $\nu_{\min} = \exp(-2)$, $\nu_{\max} = 5\,10^3$ and $e_{\max} = 1 - 10^{-6}$.

Then, the prior can be defined as convenience on $\mathbf{\Theta}_B$, 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, even if the angular measure is obtained in the limit of asymptotically dependent observations.

As usual, the prior's impact will vanish with large sample sizes, but this kind of control may be useful for small samples with prior expert knowledge regarding the amount of dependence at extreme levels.

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

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

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

$$\forall\, j \in \{1, \ldots, k\},\ \pi_{\nu,j} \propto \mathbf{1}_{[\nu_{\min}, \nu_{\max}]}\ \mathrm{logN}(m_\nu, \sigma_\nu^2). \tag{10}$$

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

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

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

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

where χ_μ is a concentration hyper parameter. Thus, for $\chi_\mu \ge 1$, 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, we define a Beta prior for e_m such that, conditionally to $(\boldsymbol{\mu}_{\cdot,1:m}, e_{1:m-1})$, the expectancy of e_m corresponds to a weight ratio p_m/ρ_{m-1} close to $1/(k - m + 1)$. Proposition 1 yields

$$e_m = \frac{p_m/\rho_{m-1}}{T_m\ (1 - p_m/\rho_{m-1})}.$$

The ideal situation $p_m/\rho_{m-1} = 1/(k-m+1)$ thus corresponds to $e_m = (T_m\,(k-m))^{-1}$, which may be greater than one. We thus set the distribution's mean to $M_{e,m} = \min\left\{(T_m(k-m))^{-1}, e_{\text{mean.max}}\right\}$, where $e_{\text{mean.max}} = 99/100$. Then, we define another concentration parameter χ_e, typically set to 1.1. Finally, we choose

$$a_m = \frac{\chi_e}{\min\{M_{e,m}, 1-M_{e,m}\}} M_{e,m}, \text{ and } b_m = \frac{\chi_e}{\min\{M_{e,m}, 1-M_{e,m}\}}(1-M_{e,m}).$$

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

The Directed acyclic graph in Figure 2 summarizes the model specification. Simulating parameters $(\boldsymbol{\mu}_{\cdot,1:k-1}, e_{1:k-1})$

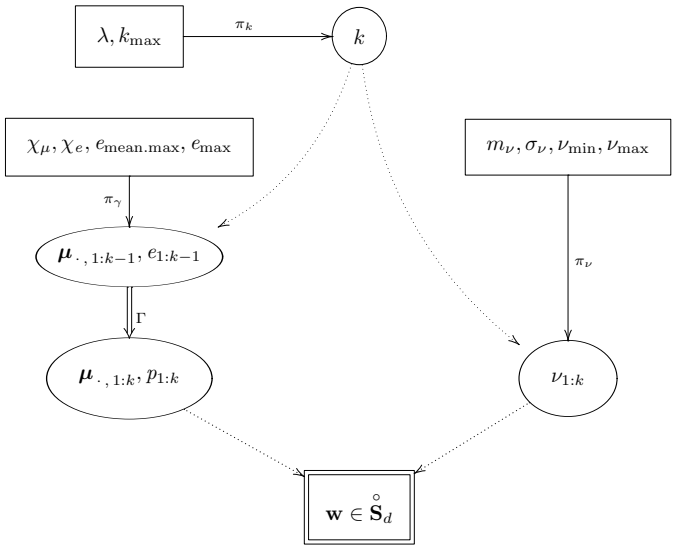

Figure 2: Representation of the conditional dependencies of the DM Bayesian model as a Directed acyclic graph. Hyper-parameters appear in simple square frames, parameters in oval frames and observations in a double square frame. Simple arrows denote probabilistic relations whereas double arrows stand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

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 $p_{1:k}$.

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 limit ourselves to weak consistency properties. 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. We show (Appendix A) that our prior on the re-parametrized parameter space induces a weakly consistent posterior at all distribution that is arbitrarily close to the model in terms of Kullback-Leibler divergence (Proposition 2 below).

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

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

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 on the densities). Let π be a prior on $\mathcal{T}$ and π_n be the posterior, given observations $\mathbf{W}_1, \ldots, \mathbf{W}_n$, which are independent, identically distributed (*i.i.d.*) according to some 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 A for details). 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 establishes 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 opens set for the Euclidean topology on the $\mathbf{\Theta}_k$'s. [3].

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* (9)*. 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. More generally, any prior obtained as a mixture of point masses and of a positive density on $\mathbf{\Theta}_B$ assigns positive mass to Euclidean open sets.

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 produce samples from the posterior, which we call *Metropolis for Dirichlet mixture*, or, in short, M-DM . It belongs to the class of *Metropolis within Gibbs* algorithms (MH-Gibbs), as described *e.g.* in Roberts and Rosenthal (2006).

[3] These open sets define the co-product topology induced by the Euclidean topology on the disjoint union $\coprod_k \mathbf{\Theta}_k$.

The key principle of the M-DM algorithm 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 starting value is generated according to a prior distribution. 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 $c_{\text{shuf}} = 1/6$.

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.

4.1. Regular moves

If $\theta_t = \left(\boldsymbol{\mu}_{\cdot,1:k-1}(t), e_{1:k-1}(t), \nu_{1:k}(t)\right) \in \boldsymbol{\Theta}_k$, 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 equal to $\log(\nu_m(t))$ 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 *re-centering* 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 C.

The acceptance probability for each regular move is classically given by (*e.g.* for e-moves affecting the m^{th} 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 C.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

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

and, for a combine move, to

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

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. In what follows, it is assumed that condition (5) holds for θ_t.

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 B 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 C.3)

$$\left|\text{Jac}(\varphi_{m_1,m_2})_{[\theta_t]}\right| = \prod_{m=1}^{k-1} \frac{\rho_{m-1}\, T_m}{(1+e_m\, T_m)^2} \prod_{m=1}^{k-1} \frac{(1+e^*_m\, T^*_m)^2}{\rho^*_{m-1}\, T^*_m}, \tag{11}$$

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-DM algorithm

Their 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 $\Theta = \boldsymbol{\Theta}_B$). 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) stating that, under regularity assumptions[4], if an aperiodic Markov chain generated by a transition kernel $K(\theta, \cdot)$ admits $\tilde{\pi}$ as an

[4]It is required 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 $\boldsymbol{\Theta}$ can be identified with a finite union of open subsets in finite dimensional euclidean spaces.

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 $\tilde{\pi}$.

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 $t \in \mathbb{N}$, $K^t(\theta_{\text{start}}, A) > 0$.

In view of section 3.3, total variation distance is more than needed, because we only know about weak consistency properties of the posterior. However, it 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, convergence in total variation implies

$$\frac{1}{T}\sum_{t=1}^{T} g(\theta_t) \underset{T\to\infty}{\longrightarrow} \mathbb{E}_{\tilde{\pi}}(g), \quad P_{\theta_{\text{start}}} \text{ almost surely}, \tag{12}$$

where $P_{\theta_{\text{start}}}$ represents the probability measure on $\left(\boldsymbol{\Theta}^{\mathbb{N}}, \mathcal{T}^{\otimes\mathbb{N}}\right)$ 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 (12).

In order to verify that (12) holds for the M-DM algorithm, we show in Appendix B the following

Proposition 3. *Let η the Lebesgue measure restricted to $\boldsymbol{\Theta}_B$. 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 (12) 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-DM algorithm. 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 (12) (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,

$$\begin{aligned} \tilde{g}(\theta) &= \int_{\mathbf{S}_d} g(\mathbf{w}) h_\theta(\mathbf{w})\ \mathrm{d}\mathbf{w} \\ &:= \langle g, h_\theta \rangle\ . \end{aligned} \tag{13}$$

The function $\tilde{g}$ is bounded by $\|g\|_\infty$, and its continuity (for the weak topology) may be verified[5]. Consequently, provided that the true measure h_0 satisfies the assumptions of Proposition 2 (so that the posterior is weakly consistent

[5]The arguments are the same as those leading to the continuity of κ, in the proof of Proposition 2

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 (12) shows that

$$\lim_{n\to\infty}\left(\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\langle g, h_{\theta_t^n}\rangle\right) = \langle g, h_0\rangle \qquad (h_0 \times P_{\theta_{\text{start}}}\text{-a.s.})\,. \tag{14}$$

where θ_t^n is the current state at time t of the algorithm, with objective probability the posterior π_n.

6. Convergence assessment

6.1. Choice of the monitored quantity

In this section, we propose a method to verify in practice that the asymptotic domain of validity (14) has approximately been reached, for a given data set and an an output of the M-DM algorithm. 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 the generated densities $(h_{\theta_t})_t$ themselves. One solution is to extract suitable numerical quantities that represent the generated densities, in relation to (14), 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 (14) can easily be derived in arbitrary dimension. Then the M-DM estimates $\frac{1}{T}\sum_{t=1}^{T}\langle g, h_{\theta_t}\rangle$ will be compared to the reference value (the true distribution for simulation 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 $\nu\ \mu_i > 1$, 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

$$\begin{aligned}
\langle g, h\rangle &= \int_{\mathbf{S}_d} g(\mathbf{w})h(\mathbf{w})\ \mathrm{d}\mathbf{w} \\
&= \frac{\Gamma(\nu)\Gamma(\tilde{\nu})}{\prod_{i=1}^{d}\Gamma(\mu_i\nu)\Gamma(\tilde{\mu}_i\tilde{\nu})}\int_{\mathbf{S}_d}\prod_{i=1}^{d} w_i^{(\mu_i\nu+\tilde{\mu}_i\tilde{\nu}-1)-1}\ \mathrm{d}\mathbf{w} \\
&= \frac{\Gamma(\nu)\Gamma(\tilde{\nu})}{\prod_{i=1}^{d}\Gamma(\mu_i\nu)\Gamma(\tilde{\mu}_i\tilde{\nu})}\frac{\prod_{i=1}^{d}\Gamma(\mu_i'\nu')}{\Gamma(\nu')} \\
&:= \mathcal{I}_{\boldsymbol{\mu},\nu}(\tilde{\boldsymbol{\mu}}, \tilde{\nu})
\end{aligned} \tag{15}$$

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 (15) 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

$$\hat{g}_n^{\text{nonP}} \ = \ \frac{1}{n}\sum_{j=1}^{n} g(\mathbf{W}_j). \tag{16}$$

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

$$\delta_n^{\text{nonP}} = \frac{1}{\sqrt{n}} \left[\text{Var}_{h_0}(g)\right]^{1/2} = \frac{1}{\sqrt{n}} \left[\mathbb{E}_{h_0}(g^2) - \left(\mathbb{E}_{h_0}(g)\right)^2\right]^{1/2}, \tag{17}$$

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

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, (14) may be reformulated in terms of convergence of the posterior predictive density in a neighborhood of the simplex point $\boldsymbol{\mu}$. In practice, choosing such a g (see Appendix D) allows to check that the posterior predictive behaves well in regions of interest (for example, in the regions where the observed angular data concentrate). Also, in this paper, the Dirichlet g's are chosen according to the angular data set: their mean vector are drawn in the neighborhoods of the data points. More details are gathered in Appendix D.

6.2. Assessing convergence in practice

For each case study, the M-DM algorithm was ran J times (typically, $J = 4$ or $J = 8$) with starting values generated from the prior. For the sake of simplicity, we use the convergence assessment tools available in R package coda. First, the stationarity of single 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 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 Gelman ratio statistic R_G (shrink factor) converges to 1 under the null-hypothesis and a typical requirement is that $R_G < 1.1$.

Beside stationarity and mixing properties, the accuracy of the estimate is of primarily interest. Suppose first that h_0 is a known Dirichlet mixture (simulation experiment). Discarding the first T_1 iteration 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-DM algorithm is

$$\hat{g}_n^{\text{DM}}(J, T_1, T_2) = \frac{1}{J(T_2 - T_1)} \sum_{j=1}^{J} \sum_{t=T_1+1}^{T_2} \langle g, h_{\theta_t(j)} \rangle$$

Each term of the summation has analytical expression derived from (15). The *exact DM error* is then

$$\Delta_n^{\text{DM}}(J, T_1, T_2) = \left| \hat{g}_n^{\text{DM}}(J, T_1, T_2) - \langle g, h_0 \rangle \right| .$$

As a summary, the error ratio

$$r^{\text{DM}} = \frac{\Delta_n^{\text{DM}}(J, T_1, T_2)}{\delta_n^{\text{nonP}}} \tag{18}$$

may be used as an indicator of the posterior mean estimate's accuracy.

If h_0 is unknown, accuracy of the fit may still be assessed by comparing the model estimate with its empirical counterpart, *i.e.* by replacing $\langle g, h_0 \rangle$ with $\hat{g}^{\text{nonP}}$ (see (16)) in (18) and δ_n^{nonP} with its estimate $\hat{\delta}_n^{\text{nonP}}$. This defines the *empirical DM error* $\hat{\Delta}_n^{\text{DM}}$ and the empirical error ratio $\hat{r}^{\text{DM}} = \frac{\hat{\Delta}_n^{\text{DM}}}{\hat{\delta}_n^{\text{nonP}}}$.

In practice, it is impossible to span the whole range of test functions and to conclude that the MC MC output *as a whole* is stationary. However, reasonable p-values for the Heidelberger tests, together with Gelman ratio close to one and moderate error ratios, on a representative family of test function, confers some credibility to the estimates constructed from the posterior samples.

7. Results

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

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 0.3 & 0.2 & 0.475 \\ 0.6 & 0.1 & 0.175 \\ 0.1 & 0.7 & 0.35 \end{pmatrix}, \qquad (19)$$
$$p_0 = (5/12, 1/4, 0.5, 1/3), \text{ and } \nu_0 = (15, 11, 20).$$

Figure 3 compares the true density with the posterior predictive resulting from one chain[6]. The Grey dots are the corresponding angular points over the simplex $\overset{\circ}{\mathbf{S}}_d$ with $d = 3$.

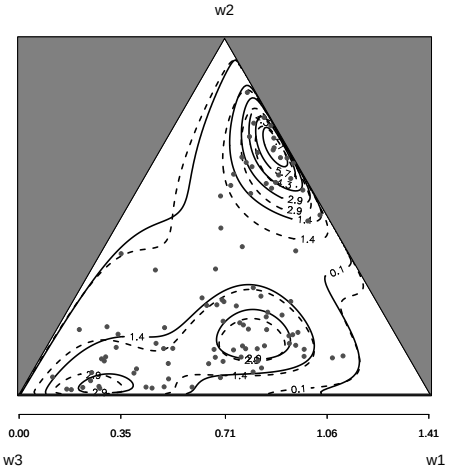

Figure 3: Predictive angular density contours (solid lines) obtained *via* the M-DM algorithm, on the two-dimensional simplex $\mathbf{S}_3$, inferred with 100 simulated points (Grey points) simulated from the true density (dotted lines) defined by (19).

The predictive angular density appears to reproduce well the characteristics of the mixture. To complement this visual check with a quantitative convergence assessment, we follow the procedure described in Section 6.2. Four parallel M-DM chains of 50 000 iterations are run (the first 10 000 are discarded). Five Dirichlet test function are randomly chosen (namely, the Dirichlet mean vectors are drawn among the angular data points, see Appendix D for details). The minimum p-values (over the five test functions) for the Heidelberger test applied to each 'mapped' chain are respectively (0.13, 0.10, 0.13, 0.06). Setting a significance level of 0.05, all the chains are deemed stationary. The multivariate Potential scale reduction factor equals $R_G = 1.01$, indicating good mixing properties. The error ratios r^{DM} (see (18)) for the five test functions are (0.02, 0.24, 0.51, 1.04, 0.48), respectively. All these ratios but one are lower than one, as required.

As a comparison, the same procedure is followed with the algorithm proposed by Boldi and Davison (2007). The predictive densities are similar to those obtained with the M-DM algorithm but convergence issues arise: for the same test functions as above, the Heidelberger p-values are $(0.015, 0.001, 0.07, 2 \cdot 10^{-6})$ so that only the third chain is not rejected. For this chain, the error ratios are (0.11, 0.63, 0.83, 1.36, 0.46), which is on average slightly more than those

[6]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.

obtained with the M-DM algorithm. Decreasing the limiting p-value for rejection to 0.01, two chains can be retained, but then the Gelman ratio is equal to 2.17.

For a more immediate convergence diagnostic, Figure 4 shows the evolution of the quantities $\langle g, h_{\theta_t(j)} \rangle$ (as defined in (13)), where $j \in \{1, \ldots, 4\}$ is the chain index, and of the mean estimates $\hat{g}^{DM}(\{j\}, 0, T) = \frac{1}{T} \sum_{t \leq T} \langle g, h_{\theta_t(j)} \rangle$., for one given test function [7]. Clearly, the mixing properties of the original algorithm are not as good as in the re-parametrized version, so that it should be ran with a much higher number of iterations for the estimates to be fully reliable with real data.

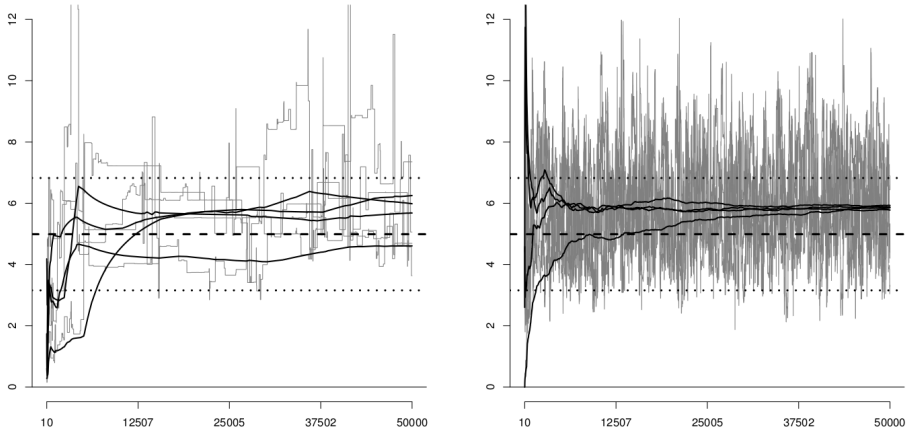

Figure 4: Convergence monitoring with three-dimensional data in the original DM model (left panel) and in the re-parametrized version (right panel), with four parallel chains in each model. Grey lines: Evolution of $\langle g, h_{\theta_t(j)} \rangle$. Black, solid lines: cumulative mean. Dashed line: true value $\langle g, h_0 \rangle$. Dotted lines: true value +/- 1 theoretical standard deviation δ_n^{nonPar} of the empirical mean estimate with $n = 100$ points.

7.2. Example: stimulated five-dimensional data

We now turn to higher dimensional problems, for which the density on the full dimensional simplex cannot be easily represented, except by projection or marginalization over triplets or pairs of coordinates. A 100- points data set is simulated from a four-components DM distribution with parameters

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

$$p_0 = (0.2, 0.1, 0.2, 0.5)\,, \text{ and } \nu_0 = (30, 40, 20, 25).$$

Four parallel chains of length $200\,10^3$ are run in each model. As in the three dimensional example, 5 Dirichlet test functions are randomly chosen to monitor convergence. As an example, for one test function g with Dirichlet parameter $\alpha = \nu\,.\,\mu \simeq (2.3, 1.1, 1.7, 5.9, 13.9)$, the evolution of the $\langle g, h_{\theta_t(j)} \rangle$ and $\hat{g}^{DM}(\{j\}, 0, T)$ is shown in Figure 5. Table 1 gathers the results of the convergence diagnostics performed after randomly selecting five test functions, as in the three-dimensional case.

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-DM algorithm. The computational burden remains reasonable: the typical run-time is of five minutes for one chain. One practical implication of the slow mixing on the

[7] The Dirichlet test function is the third of the list, *i.e.* a Dirichlet density with parameter $\nu\boldsymbol{\mu} \simeq (17.7,\ 25.2,\ 1.001)$

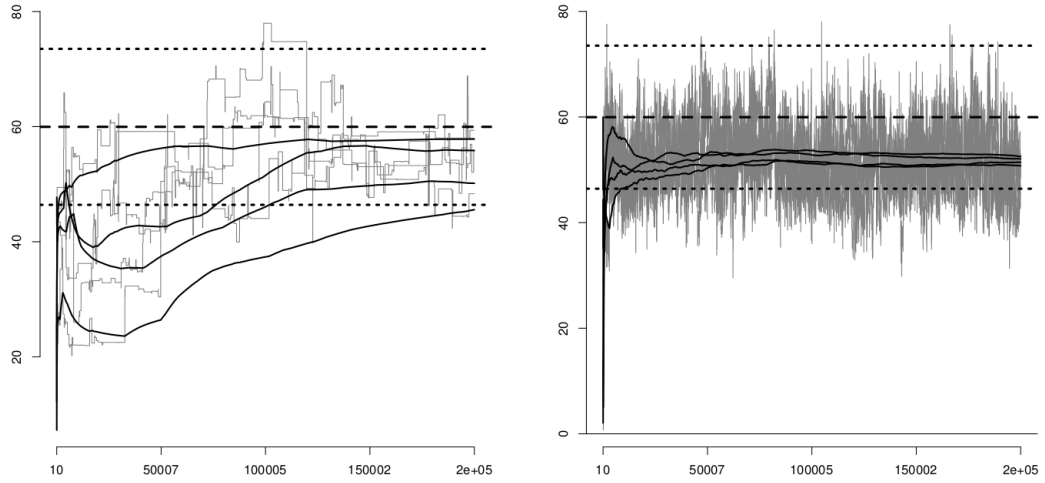

Figure 5: Convergence monitoring with five-dimensional data in the original DM model (left panel) and in the re-parametrized version (right panel), with four parallel chains in each model. Grey lines: Evolution of $\langle g, h_{\theta_t(j)} \rangle$. Black, solid lines: cumulative mean. Dashed line: true value $\langle g, h_0 \rangle$. Dotted lines: true value +/- 1 theoretical standard deviation δ_n^{nonPar} of the empirical mean estimate with $n = 100$ points.

	#{stationary}	HW_1	HW_2	HW_3	HW_4	R_G	r_1	r_2	r_3	r_4	r_5
M-DM	3	0.05	0.06	0.01	0.07	1.07	0.27	0.65	0.03	0.17	0.04
BD	1	0.01	0.07	2.10^{-5}	0.03	'NA'	0.45	0.45	0.18	0.42	0.81

Table 1: Simulated five dimensional data: convergence statistics for the output of the M-DM algorithm (first line) and the original version from Boldi and Davison (2007) (second line). First column: number of chains retained by the Heidelberger and Welches test. Columns 2-5 :minimum p-values (over the five test functions) of the Heidelberger and Welches' statistics applied to each chain. Column 6: Gelman ratio. Columns 7-19: accuracy of the estimate: ratio statistics defined by (18) for five test functions.

original parametrization is that posterior credible intervals are difficult to estimate. As an example, Figure 6 displays, for the two parametrizations, the estimated posterior mean of the bi-variate angular density for the coordinates pair $(2, 5)$, 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 a chain that which stationarity was not rejected (for a 0.05 p-value) by the Heidelberger test. 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 $(0, 1)$.

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). 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). Eight chains of 10^6 iterations each were generated. Discarding half of the iterations and setting the minimum p-value to 0.01, 4 (*resp.* 5 chains) cannot be rejected by the stationarity test with the re-parametrized algorithm (*resp.* with the original one). For those chains, the Gelman ratio is 1.08 (*resp.* 1.75).

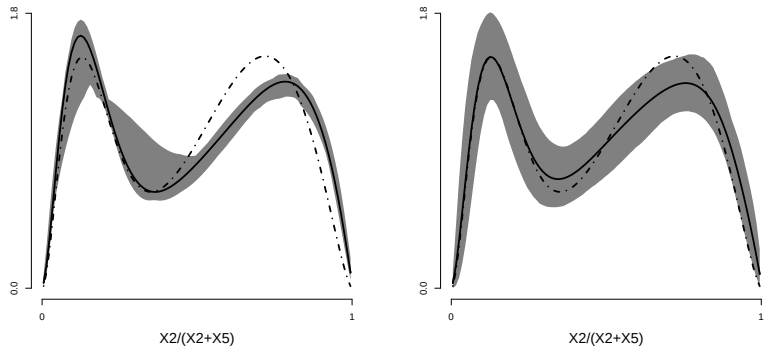

Figure 6: Simulated five-dimensional data (100 points): Bi-variate angular posterior predictive densities for the pair $(2,5)$. Left panel: Original algorithm; Right panel: re-parametrized version. Dash-dotted line: true density; solid line: posterior predictive; Grey area: posterior credible set at level 0.9.

This indicates again that mixing remains acceptable in the re- parametrized DM model, provided the run length is long enough, which is not the case in the original version. Figure 7 shows the projection of the predictive density on the three two-dimensional simplex faces. Again, the mean estimates obtained with the original MC MC algorithm are very similar but the posterior $0.05-0.95$ quantiles are thinner (not shown). This example allows to verify that our estimates (after fitting the model on the five dimensional data set) are close to those found by Boldi and Davison (2007) using an EM algorithm, for the considered coordinate pairs.

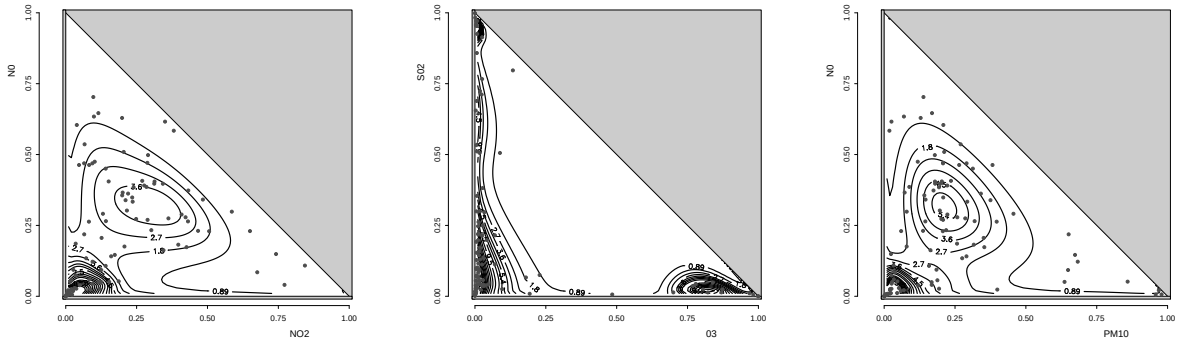

Figure 7: Five dimensional Leeds data set: posterior predictive density. Black lines: projections of the predictive angular density defined on the four-dimensional simplex $\mathbf{S}_5$ onto the two-dimensional faces. Grey dots: projections of the 100 points with greatest L^1 norm.

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 $\lambda, \sigma_\nu, \chi_\mu, \chi_e$, which correspond respectively to the mean parameter for k, the variance (on the log scale) of the shape parameters, the concentration of the mean vectors around the 'current' partial centers of mass, and the concentration of the eccentricity parameters around the objective value (see section 3.2). Also, we verify that defining the prior distribution of $(\boldsymbol{\mu}, \mathbf{e})$ jointly, as in section 7.2, 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 such 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 made to vary as in the preceding setting.

The default hyper-parameter values are set to

$$\lambda = 5\,,\ k_{\max} = 15\,,$$
$$m_\nu = \log(60)\,,\quad \sigma_\nu^2 = \log(1+5^2)\,,\quad \log(\nu_{\min}) = -2\,,\quad \log(\nu_{\max}) = 5000\,,$$
$$\chi_e = 1.1\,,\quad e_{\text{mean.max}} = 0.99\quad e_{\max} = 1 - 10^{-6}$$
$$\chi_\mu = 1.1\,.$$

Starting from this, the hyper-parameters $\lambda, \sigma_\nu, \chi_\mu, \chi_e$ are perturbed, one at a time. Namely, the model is successively fitted with $\lambda \in \{1, 3, 5, 7, 10, 12\}$ (the other parameters being set to the default except for $\lambda \in \{10, 12\}$ for which $k_{\max} = 20$), $\sigma_\nu^2 \in \{\log(1+(0.5)^2), \log(1+1^2), \log(1+2^2), \log(1+5^2), \log(1+10^2), \log(1+20^2)\}$, $\chi_\mu \in \{0.5, 1, 1.5, 2, 4, 8\}$, $\chi_e \in \{0.5, 1, 1.1, 1.5, 3, 6\}$. For each hyper-parameters value, four chains are ran in parallel, with a burn-in period of 100×10^3 followed by another period 100×10^3 iterations. Five Dirichlet test functions are chosen and the quality of the fit is assessed in terms of multivariate Gelman ratio and of the average error ratio r^{DM} over the five test functions. Figure 8. On both panels, lower values indicate better properties. It appears that the hyper-parameter λ ruling the number of components has only a limited impact, except that setting $\lambda = 1$ affects the quality of the estimates. Increasing values of λ does not alter the quality of the fit. When $\boldsymbol{\mu}$ and e are *a priori* dependent, as in section 3.2, the hyper-parameter σ_ν on the variance of the shape parameters has only a limited impact: the scores are approximately constant over the six values. As for the concentration hyper-parameter χ_e and $\chi_{\boldsymbol{\mu}}$, only very large or low values affect the quality of the fit and the mixing properties. For both of them, results are indistinguishable on the range $[1, 3]$. 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 fit is more sensitive to σ_ν and the mixing properties are generally poor. This result is somewhat in adequacy with the findings of Boldi and Davison (2007), who concluded (for the original model) that the prior on the shape should be defined as flat as possible. To conclude this part, the structure of the prior defined in section 3.2 appears to be relatively robust to the hyper-parameters specification, compared to the simplified version where $\boldsymbol{\mu}$ and $\mathbf{e}$ are *a priori* independent.

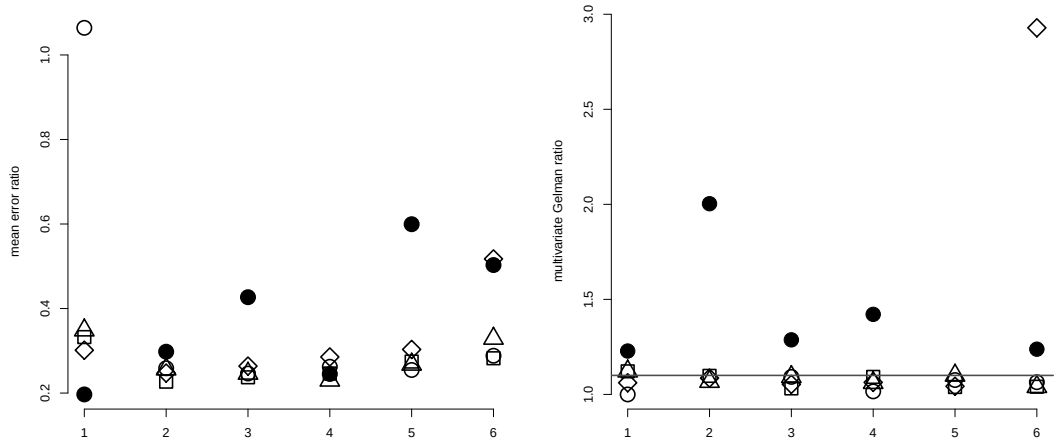

Figure 8: Influence of the prior specification on the quality of the fit (left panel) and on the chains' mixing property (right panel). •, simplified prior on $(\boldsymbol{\mu}, e)$, influence of σ_ν (variance of the shapes); □, dependent prior on $(\boldsymbol{\mu}, e)$, influence of σ_ν; ∘, influence of λ (mean number of mixture components); ⋄, influence of χ_μ (concentration of mean vectors); △, influence of χ_e (concentration of eccentricities); Grey line, level 1.1 for the Gelman ratio.

7.5. Comparison with other methods for bi-variate data

Here, the M-DM algorithm 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 model, 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).

In this study, 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). we respectively denote these 'true' distributions $H_{0,L}, H_{0,AL}, H_{0,DM}$. Contrary to the two other ones, the asymmetric logistic distribution has point masses at 0 and 1. For each $H_{0,m}$ ($m \in \{L, AL, DM\}$), 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, the bi-variate threshold (u_1, u_2) is set 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 $[u_1, \infty) \times [u_2, \infty)$. The number of MC MC 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. 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 $\hat{H}$.

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 $\hat{H} - H_i$ are plotted. In this bi-variate setting, 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 estimates. At first view, the possibility for point masses on 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.

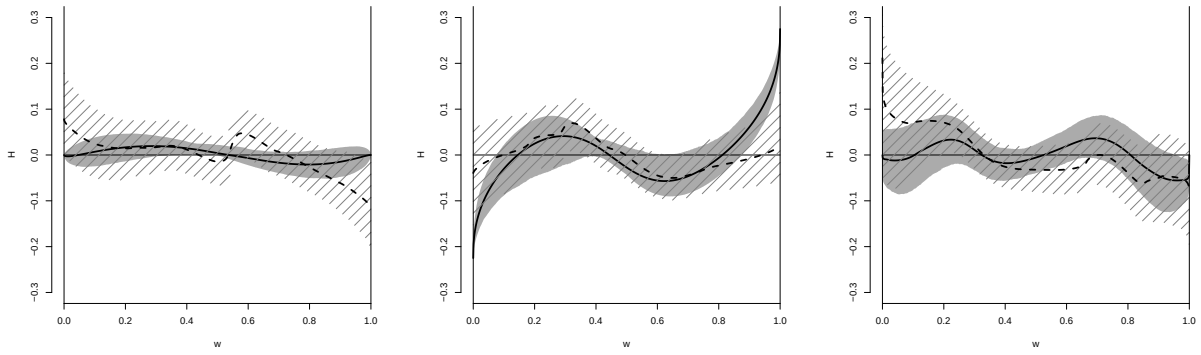

Figure 9: Error of the predictive angular *cdf* (solid lines) on the segment [0,1]. From left to right: data from a Logistic, an Asymmetric logistic and from a DM distribution. Solid line and Grey area: Dirichlet Mixture mean estimate and $0.1 - 0.9$ posterior quantiles; dashed line and dashed area: *idem* in the non-parametric model.

For a more quantitative assessment, the performance of the posterior mean estimates $\hat{H}$ for a given 'true' H_0 are compared in terms of mean integrated squared error loss (MISE), which is $\text{MISE}(\hat{H}, H_0) = \int_0^1 [\hat{H}(w) - H_0(w)]^2 \, \mathrm{d}w$, and the scores are averaged over the 100 data sets, for each underlying distribution. Table 2 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 MC MC steps is large enough.

Table 2: 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. The required computational effort is somewhat low; 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 accurate results, and the computational time would naturally increase. Still, the possibility to handle the five dimensional case may open the road to new modeling approaches in environmental applications: Consider for example five adjacent cells on the grid of a spatial climate model (see Figure 10). Environmental variables such as temperature or precipitation observed

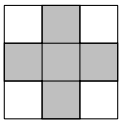

Figure 10: Five adjacent cells (Grey squares) on a two dimensional discretization grid

simultaneously on the five adjacent locations are likely to exhibit some dependence at asymptotic levels and we believe that the stability properties of the Dirichlet mixtures under conditioning and marginalization could be exploited in the context of extreme events prediction or weather generators.

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. Proofs

Appendix A.1. 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, $C_m = \{i \in \{1,\dots,d\} \ : \ \gamma_{i,m-1} - \mu_{i,m} < 0\}$. Thus, for $i \notin C_m$, the map $t \mapsto \gamma_{i,m} + t(\gamma_{i,m} - \mu_{i,m})$ is non decreasing. Thus,

$$\forall i \notin C_m, \forall t > 0, \gamma_{i,m} + t\,(\gamma_{i,m} - \mu_{i,m}) > 0\,,$$

whence

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

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 $\theta = (\boldsymbol{\mu}_{\cdot,1:k-1}, e_{1:k-1}, \nu_{1:k}) \in \boldsymbol{\Theta}_k$. 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 A.2. Weak consistency of the posterior

For convenience, we restate below Schwartz's theorem (Schwartz, 1965, Theorem 6.1, *p.*22). The proof 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 $\mathbf{S}_d = \{(w_1,\dots,w_{d-1}) : w_i \geq 0, \sum_1^{d-1} w_i \leq 1\}$.

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 $\mathcal{T}$. For us, $\Theta = \boldsymbol{\Theta}_B$ defined in (9) and $\mathcal{T}$ is the Borel σ-field associated with the topology induced by the Euclidean topology on the co-product space Θ_B. Also, 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 set of DM distributions satisfying (2), indexed by $\boldsymbol{\Theta}_B$.

As for random variables, the infinite sequence $(\mathbf{W})_\infty = \left\{\mathbf{W}_j,\ j \geq 0\right\}$ 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

$$\mathcal{H}_0 :\ h = h_0 \text{ versus } \mathcal{H}_1 :\ h \in U^c,$$

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 does not require identifiability, and which we restate for convenience. A self contained proof 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

$$\pi_n(U \cap \mathcal{M}_\Theta) \underset{n\to\infty}{\longrightarrow} 1\ ,\ h_0\textit{-almost surely.} \tag{A.1}$$

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.

Let $\mathcal{M}$ be endowed with the Borelian σ-field $\mathcal{B}(\mathcal{M})$ relative to the weak topology on $\mathcal{M}$. It is easily verified that the intersections of open sets in $\mathcal{M}$ with $\mathcal{M}_\Theta$ are measurable for the DM model (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 $\tilde{\pi}(U) = \pi(U \cap \mathcal{M}_\Theta)$. For the sake of simplicity, the $\tilde{}$ will be omitted and π denotes 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 $\mathcal{M}$, it is easy to find a uniformly consistent sequence of tests for $\mathcal{H}_0$ versus $\mathcal{H}_1$. 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 alone 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 base and we can exhibit a set $\Omega_0 \subset \mathcal{S}^{\mathbb{N}}$, with $h_0(\Omega_0) = 1$, on which convergence (A.1) occurs for *all* neighborhood 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 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

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

(Note that, by compacity of $\bar{B}_\epsilon$, the *extrema* are reached, which ensures positivity of the *infima*)

Hence, $\forall (\boldsymbol{\mu}, \mathbf{e}, \boldsymbol{\nu}) \in \bar{B}_\epsilon$, $\forall \mathbf{w} \in \overset{\circ}{\mathbf{S}}_d$, $\forall m \leq k$,

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

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

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

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

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

Let $C_1 = \max\{|D_1|, |D_2|\}$ and $C_2 = \max\{|a_{\min} - 1|, |a_{\max} - 1|\}$. 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 B. 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 $\theta = (\boldsymbol{\mu}, e, v) \in \mathbf{\Theta}_{Bk}$, 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 $v_{m-2k_{\max}+2}$ (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, denoting $\theta^0 = \theta_{\text{start}}$, 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, we have

$$K^{3k-2}(\theta, A|c_{1:3k-2}) \geq \int_{A_1} \cdots \int_{A_{3k-2}} \prod_{t=1}^{3k-2} q_t^k \, r_t^k \left(\theta_{t-1}, \tilde{\theta}(\theta_{t-1}, x^t)\right) \mathrm{d}x^1 \cdots \mathrm{d}x^{3k-2} ,$$

where, $\forall 1 \leq m \leq 3k-2$, $\theta^m \in A_1 \times \cdots \times A_m \times E_{m+1}^k \times \cdots \times E_{3k-2}^k$.

Further, for $\theta^0, x^1, \ldots, x^{3k-2} \in \mathbf{\Theta}_{Bk} \times A_1 \times \cdots \times A_{3k-2}$, we have

- $\tilde{\pi}(\theta^0) > 0$
- $\forall m \geq 1,\ q_m(\theta^{m-1}, \theta^m) > 0$ and $q_m(\theta^m, \theta^{m-1}) > 0$

So that each term of the integrand is positive, and we have $K^{3k-2}(\theta_{\text{start}}, A|c_{1:3k-2}) > 0$. Thus, $K^{3k-2}(\theta_{\text{start}}, A) > 0$.

Assume now that $\theta_{\text{start}} \notin \mathbf{\Theta}_{Bk}$. In such a case, the probability of proposing and accepting trans-dimensional moves until the chain reaches $\mathbf{\Theta}_{Bk}$ is positive. Consequently, $\mathbf{\Theta}_{Bk}$ is accessible from θ_{start}, which completes the proof.

Invariance of the posterior distribution under the M-DM kernel

Since the whole M-DM kernel K is a weighted average of partial kernels defined in section 4, it is enough to show that the posterior is invariant under each of them. The invariance under trans-dimensional moves is ensured by the fact that the acceptance ratios r_{split} and r_{combine} defined in section 4.2 satisfy Green (1995)'s balance condition. Also, each 'regular' kernel $K_m(\theta, \cdot)$ (*i.e.* affecting one $\boldsymbol{\mu}_{\cdot,m}$, one ν_m or one e_m corresponds to a *Metropolis-within-Gibbs* partial kernel, as defined *e.g.* in Roberts and Rosenthal (2006), so that, if we denote $\text{Im}(K_m, \theta) \subset \mathbf{\Theta}$ the image of $K_m(\theta, \cdot)$, η_m the reference Lebesgue measure on $\text{Im}(K_m, \theta)$, q_m the proposal density (w.r.t. η_m) and r_m acceptance probability, then, following Roberts and Rosenthal (2006, Section 4), the so-called *balance equation*, $\tilde{\pi}(\theta) q_m(\theta, \theta^*) r_m(\theta, \theta^*) = \tilde{\pi}(\theta^*) q_m(\theta^*, \theta) r_m(\theta^*, \theta)$, ensures the invariance of $\tilde{\pi}$ under K_m.

We only need to show the invariance under the shuffle moves.

Let $r(\theta) = r_{\text{shuffle},m_1,m_2}(\theta, \theta^*)$ denote the acceptance probability of the *shuffle* move as described in Section 4.3, for a transposition φ_{m_1,m_2}, so that $\theta^* = \varphi_{m_1,m_2}(\theta) := \varphi(\theta)$. Let K_s be the corresponding transition kernel (*i.e.*, the transition kernel conditionally to proposing a shuffle move affecting m_1 and m_2). We derive a sufficient condition on r for the posterior distribution π_n to be invariant under K_s. The proposal kernel Q_s, conditionally to the acceptance of the shuffle move, is the point mass $Q_s(\theta, A) = \delta_{\varphi(\theta)}(A) = \mathbf{1}_A(\varphi(\theta))$, for $A \subset \mathbf{\Theta}_B$. The shuffle kernel K_s may thus be written as

$$\begin{aligned} K_s(\theta, A) &= r(\theta)\mathbf{1}_A(\varphi(\theta)) + (1 - r(\theta))\mathbf{1}_A(\theta) \\ &= r(\theta)\mathbf{1}_{\varphi^{-1}(A)}(\theta) + (1 - r(\theta))\mathbf{1}_A(\theta)\ . \end{aligned}$$

and the shifted measure of A is

$$\begin{aligned} K_s.\pi_n(A) &= \int_{\varphi^{-1}(A)} \pi_n(\theta) r(\theta)\ \mathrm{d}\theta + \int_A (1 - r(\theta))\pi_n(\theta)\ \mathrm{d}\theta \\ &= \int_A \pi_n\left(\varphi^{-1}(\theta^*)\right) r(\varphi^{-1}(\theta^*))\, |\text{Jac}(\varphi)|^{-1}_{[\varphi^{-1}(\theta^*)]}\ \mathrm{d}\theta^* + \\ &\quad \ldots \int_A (1 - r(\theta))\pi_n(\theta)\ \mathrm{d}\theta \\ &= \pi_n(A) + \\ &\quad \ldots \int_A \pi_n\left(\varphi^{-1}(\theta^*)\right) r(\varphi^{-1}(\theta^*))\, |\text{Jac}(\varphi)|^{-1}_{[\varphi^{-1}(\theta^*)]} - r(\theta^*)\pi_n(\theta^*)\ \mathrm{d}\theta^*\ . \end{aligned}$$

A sufficient condition for $K_s.\pi_n(A) = \pi_n(A)$ is thus $\pi_n(\theta)\, r(\theta)\, |\text{Jac}(\varphi)|^{-1}_{[\theta]} = r(\theta^*)\pi_n(\theta^*)$, or

$$\forall \theta \in \mathbf{\Theta}_B,\ \frac{r(\theta)}{r(\theta^*)} = \frac{\pi_n(\theta^*)}{\pi_n(\theta)}\, |\text{Jac}(\varphi)|_{[\theta]} \tag{B.1}$$

Now, since φ is the transposition of two components of the Ψ-parametrization, we have $\varphi = \varphi^{-1}$, and

$$|\text{Jac}(\varphi)|_{[\theta]} = \sqrt{|\text{Jac}(\varphi)|_{[\theta]}\ |\text{Jac}(\varphi)|_{[\theta]}} = \sqrt{\frac{|\text{Jac}(\varphi)|_{[\theta]}}{\left|\text{Jac}(\varphi^{-1})\right|_{[\varphi(\theta)]}}} = \sqrt{\frac{|\text{Jac}(\varphi)|_{[\theta]}}{|\text{Jac}(\varphi)|_{[\theta^*]}}}\ ,$$

so that (B.1) holds if we set $r(\theta)$ to

$$r(\theta) = \min\left(1, \frac{\pi_n(\theta^*)}{\pi_n(\theta)}\, |\text{Jac}(\varphi)|_{[\theta]}\right)$$

It must be noted that the above argument is not valid for general permutations of indices $\varphi_{m1,\ldots,m_d}$, unless the condition $\varphi = \varphi^{-1}$ holds.

□

Appendix C. M-DM algorithm details

Appendix C.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 $\mathbf{W}_{1:n} = (\mathbf{W}_1, \dots, \mathbf{W}_n)$:

$$q_\mu(\boldsymbol{\mu}_{\cdot,m}(t), \cdot) = \sum_{j=1}^{n} \tilde{p}_j \ \mathrm{diri}(\cdot \,|\, \tilde{\boldsymbol{\mu}}_{\mathbf{W}_j}, \tilde{\nu}).$$

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 $\boldsymbol{\gamma}_0 = (1/d, \dots, 1/d)$. is the centroid of the simplex. The concentration parameter is set to $\tilde{\nu} = \frac{d}{\tilde{\epsilon}_w}$, So that each component $\mathrm{diri}(\cdot \,|\, \tilde{\boldsymbol{\mu}}_{\mathbf{W}_j}, \tilde{\nu})$ is bounded, with mode at $\mathbf{W}_j$.

Finally, the weights $(\tilde{p}_1, \dots, \tilde{p}_n)$ 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

$$\tilde{\tilde{p}}_j = \mathrm{diri}(\mathbf{W}_j \,|\, \tilde{\boldsymbol{\mu}}_\mu, \tilde{\nu}_\mu).$$

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 C.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

$$q_{\mu,\mathrm{split}}(\theta_t, \cdot) = \sum_{j=1}^{n} \tilde{p}_j^{\mathrm{split}} \ \mathrm{diri}(\cdot \,|\, \tilde{\boldsymbol{\mu}}_{\mathbf{W}_j}, \tilde{\nu})$$

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 $\tilde{p}_j^{\mathrm{split}}$ 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 $\tilde{\epsilon}_\mu^{\mathrm{split}} = 0.5$ 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 k^{th} eccentricity parameter e^*_k is generated, conditionally to the proposed mean vector $\boldsymbol{\mu}^*_{\cdot,k}$, according to the prior distribution:

$$q_{e,\mathrm{split}}(\theta_t, \cdot \,|\boldsymbol{\mu}^*_{\cdot,k}) = \pi_{e,k}(\cdot \,|\boldsymbol{\mu}_{\cdot,1:k-1}, \boldsymbol{\mu}^*_{\cdot,k}, e_{1:k}) \,.$$

Finally, the last shape parameter ν^*_{k+1} is generated according to the proposal distribution for regular ν-moves, conditionally on $\nu_k(t)$:

$$q_{\nu,\mathrm{split}}(\theta_t, \cdot) = q_\nu(\nu_k(t), \cdot).$$

Appendix C.3. Jacobian term in the acceptance ratio for shuffle moves

Here is derived the closed form of $\mathrm{Jac}(\varphi)$ appearing in (11).

The indices m_1, m_2, and we denote G the local diffeomorphism deduced from Γ:

$$G : \boldsymbol{\Theta}_{Bk} \subset \mathbb{R}^{3k-2} \longrightarrow G(\boldsymbol{\Theta}_{Bk}) \subset \mathbb{R}^{3k-2}$$
$$\left(\boldsymbol{\mu}_{\cdot,1:k-1}, e_{1:k-1}, \nu_{1:k}\right) \longmapsto \left(\boldsymbol{\mu}_{\cdot,1:k-1}, p_{1:k-1}, \nu_{1:k}\right) .$$

Recall that $\varphi(\theta) = \Gamma^{-1} \circ \tau \circ \Gamma(\theta)$, so that

$$\mathrm{Jac}(\varphi)_\theta = \mathrm{Jac}(G^{-1})_{\tau\circ\Gamma(\theta)} \, \mathrm{Jac}(\tau)_{\Gamma(\theta)} \, \mathrm{Jac}(G)_\theta \,.$$

The determinant of a transposition is -1, so that

$$|\mathrm{Jac}(\varphi)_\theta| = \left|\frac{\mathrm{Jac}(G)_\theta}{\mathrm{Jac}(G)_{\theta^*}}\right|,$$

and we only need to compute $\mathrm{Jac}(G)$. 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, $\mathrm{Jac}(G) = \left|M_{p,e}\right|$.

Since p_m depends only on the $\{\boldsymbol{\mu}_{\cdot,j}, e_j : j \le m\}$, we have

$$\left|M_{p,e}\right| = \begin{vmatrix} \frac{\partial p_1}{\partial e_1} & 0 & \cdots & 0 \\ * & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \cdots & * & \frac{\partial p_{k-1}}{\partial e_{k-1}} \end{vmatrix}$$

whence

$$\left|M_{p,e}\right| = \prod_{m=1}^{k-1} \frac{\partial p_m}{\partial e_m} \quad .$$

From Proposition 1, we have

$$\begin{aligned} \frac{\partial p_m}{\partial e_m} &= \frac{\partial}{\partial e_m}\left(\rho_{m-1}\frac{e_m T_m}{1+e_m T_m}\right) \\ &= \frac{\rho_{m-1}\, T_m}{(1+e_m T_m)^2} \quad . \end{aligned}$$

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

$$|\mathrm{Jac}(\varphi)| = \prod_{m=1}^{k-1} \frac{\rho_{m-1}\, T_m}{(1+e_m T_m)^2} \prod_{m=1}^{k-1} \frac{(1+e_m^* T_m^*)^2}{\rho_{m-1}^* T_m^*} \tag{C.1}$$

where the $e_m^*, \rho_{m-1}^*, T_m^*$ are relative to the proposal parameter $\theta^* = \varphi(\theta)$.

Appendix D. Convergence assessment: Random choice of Dirichlet test functions

Let $\mathbf{W}_{1:n} = (\mathbf{W}_1, \ldots, \mathbf{W}_n)$ be an angular data set, $L \in \mathbb{N}^*$ and

$$\{g_\ell = \mathrm{diri}(\cdot\,|\tilde{\boldsymbol{\mu}}_\ell, \tilde{\nu}_\ell),\ 1 \le \ell \le L\},$$

a set of Dirichlet test functions to be constructed. 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 candidate angular data points as follows:

We fix a maximum shape parameter $\tilde{\nu}_{\max} = 300 * d$, where d is the dimension of the sample space. Then, we retain the n' angular points $\mathbf{W}_j$ which verify

$$\min_{1\le i\le d}\{W_{i,j}\} > \tilde{\nu}_{\max}$$

Then, L angular data points $\left\{\mathbf{W}_{j_\ell}, 1 \le \ell \le L\right\}$, are chosen (uniformly) among the n' candidate data points, and we set $\tilde{\boldsymbol{\mu}}_\ell = \mathbf{W}_{j_\ell}$. Finally, we fix a minimum value $\tilde{\nu}_{\min} = 5 * d$ for the test's shape parameter, as well as a multiplying constant $\chi_{\text{test}} = 1.001$, then the ℓ^{th} shape parameter is set to

$$\tilde{\nu}_\ell = \max\left\{\frac{\chi_{\text{test}}}{\min_{1\le i\le d}\tilde{\mu}_{i,\ell}},\ \tilde{\nu}_{\min}\right\}.$$

Appendix E. 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 $g = \text{diri}(\cdot\,|\,\tilde{\boldsymbol{\mu}}, \tilde{\nu})$ 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 (15).

To compute $\mathbb{E}_\theta(g^2)$, we note that

$$g^2(\cdot) = C_{\tilde{\boldsymbol{\mu}},\tilde{\nu}}\,\text{diri}(\cdot\,|\,\boldsymbol{\mu}', \nu'),$$

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 (17) follows:

$$\delta_n^{\text{nonP}} = n^{-1/2}\left[C_{\tilde{\boldsymbol{\mu}},\tilde{\nu}}\sum_{m=1}^{k} p_m\,\mathcal{I}_{\boldsymbol{\mu}_{\cdot,m}, \nu_m}(\boldsymbol{\mu}', \nu') - \left(\sum_{m=1}^{k} p_m\,\mathcal{I}_{\boldsymbol{\mu}_{\cdot,m}, \nu_m}(\tilde{\boldsymbol{\mu}}, \tilde{\nu})\right)^2\right]^{1/2}$$

Appendix F. Bi-variate distributions used in the simulation study

Appendix G. Correction on the prior proposed by 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:

$$F_\mu(\boldsymbol{\mu}_{\cdot,1}, \ldots, \boldsymbol{\mu}_{\cdot,k}) = f_{1,1}(\mu_{1,1})\, f_{1,2}(\mu_{1,2}|\mu_{1,1}) \cdots f_{1,k-1}(\mu_{1,k-1}|\mu_{1,1:k-2}) \cdots f_{d-1,k-1}(\mu_{d-1,k-1}|\mu_{1:d-1,1:k-2}),$$

where $f_{i,j}$ is a uniform distribution on the largest interval $I_{i,j}$ ($i \le d-1$, $j \le k-1$) 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 $S_{i,m} = \sum_{\ell<i}\mu_{\ell,m}$ ($1 \le m \le k$).

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 for her valuable advice.

References

Asmussen, S., 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., Teugels, J., 2004. Statistics of extremes: Theory and applications. John Wiley & Sons: New York.

Billingsley, P., 1999. Convergence of probability measures. volume 316. Wiley-Interscience.

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

Bunke, O., Milhaud, X., 1998. Asymptotic behavior of bayes estimates under possibly incorrect models. The Annals of Statistics 26, 617–644.

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

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

Einmahl, J., Segers, J., 2009. Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. The Annals of Statistics 37, 2953–2989.

Freedman, D., 1963. On the asymptotic behavior of bayes' estimates in the discrete case. The Annals of Mathematical Statistics 34, 1386–1403.

Gelman, A., Rubin, D., 1992. Inference from iterative simulation using multiple sequences. Statistical science , 457–472.

Ghosal, S., Ghosh, J., Ramamoorthi, R., 1999. Consistency issues in bayesian nonparametrics. STATISTICS TEXTBOOKS AND MONOGRAPHS 158, 639–668.

Ghosh, J., Ramamoorthi, R., 2003. Bayesian nonparametrics. Springer.

Green, P., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711.

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

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

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

Heidelberger, P., Welch, P., 1983. Simulation run length control in the presence of an initial transient. Operations Research , 1109–1144.

Meyn, S., Tweedie, R., 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., Rosenthal, J., 2004. General state space markov chains and mcmc algorithms. Probability Surveys 1, 20–71.

Roberts, G., Rosenthal, J., 2006. Harris recurrence of metropolis-within-gibbs and trans-dimensional markov chains. The Annals of Applied Probability 16, 2123–2139.

Roberts, G., Smith, A., 1994. Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. Stochastic processes and their applications 49, 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., Fougères, A.L., In press. Bayesian model averaging for multivariate extremes. Extremes .

Schwartz, L., 1965. On bayes procedures. Probability Theory and Related Fields 4, 10–26. 10.1007/BF00535479.

Tierney, L., 1994. Markov chains for exploring posterior distributions. the Annals of Statistics , 1701–1728.

Walker, S., 2004. Modern bayesian asymptotics. Statistical Science , 111–117.