Dirichlet process mixtures under affine transformations of the data

Julyan Arbel, Riccardo Corradin, Bernardo Nipoti

Dirichlet process mixtures under affine transformations of the data

Julyan Arbel

Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

Riccardo Corradin

Department of Statistics and Quantitative Methods, University of Milano Bicocca, Italy

Bernardo Nipoti

School of Computer Science and Statistics, Trinity College, Dublin, Ireland

Abstract

Location-scale Dirichlet process mixtures of Gaussians (DPM-G) have proved extremely useful in dealing with density estimation and clustering problems in a wide range of domains. Motivated by an astronomical application, in this work we address the robustness of DPM-G models to affine transformations of the data, a natural requirement for any sensible statistical method for density estimation. First, we devise a coherent prior specification of the model which makes posterior inference invariant with respect to affine transformation of the data. Second, we formalise the notion of asymptotic robustness under data transformation and show that mild assumptions on the true data generating process are sufficient to ensure that DPM-G models feature such a property. Our investigation is supported by an extensive simulation study and illustrated by the analysis of an astronomical dataset consisting of physical measurements of stars in the field of the globular cluster NGC 2419.

Keywords: Affine data transformations, Astronomical data, Asymptotics, Bayesian nonparametrics, Dirichlet process mixture models, Clustering, Multivariate density estimation.

1. Introduction

A natural requirement for statistical methods for density estimation and clustering is for them to be robust under affine transformations of the data. Such a desideratum is exacerbated in multivariate problems where data components are incommensurable, that is not measured in the same physical unit, and for which, thus, the definition of a metric on the sample space requires the specification of constants relating units along different axes. As an illustrative example, consider astronomical data consisting of position and velocity of stars,

thus living in the so-called phase-space: a metric on such a space can be defined by setting a dimensional constant to relate positions and velocities. In this setting, any sensible statistical procedure should be robust with respect to the specification of such a constant (Ascasibar and Binney, 2005; Maciejewski et al., 2009). This is specially important considering that often scarce to no a priori guidance about dimensional constants might be available, thus making the model calibration a daunting task. The motivating example of this work comes indeed from astronomy, the dataset we consider consisting of measurements on a set of 139 stars, possibly belonging to a globular cluster called NGC 2419 (Ibata et al., 2011). Globular clusters are sets of stars orbiting some galactic center. The NGC 2419, showed in Figure 1, is one of the furthest known globular clusters in the Milky Way. For each star we observe a four-dimensional

Figure 1: An image of the remote Milky Way globular cluster NGC 2419 (about 300 000 light years away from the solar system). Picture by Bob Franke, with permission (www.bf-astro.com).

vector $(Y_1, Y_2, V, [\text{Fe/H}])$, where (Y_1, Y_2) is a two-dimensional projection on the plane of the sky of the position of the star, V is its line of sight velocity and [Fe/H] its metallicity, a measure of the abundance of iron relative to hydrogen. Out of these four components, only Y_1 and Y_2 are measured in the same physical unit, while dimensional constants need to be specified in order to relate position, velocity and metallicity. A key question arising with these data consists in identifying the stars that, among the 139 observed, can be rightfully considered as belonging to NGC 2419: a correct classification would be pivotal in the study of the globular cluster dynamics. Astronomers expect the large majority of the observed stars to belong to the cluster: the remaining ones, called field stars or contaminants, are Milky Way stars, unrelated to the cluster, that happen to appear projected in the same region of the plane of the sky. In general the contaminants have different kinematic and chemical properties with respect to the cluster members. Considering the nature of the problem, this research ques-

tion can be formalised as an unsupervised classification problem, the goal being the identification of the stars which belong to the largest cluster, which can be interpreted as the NGC 2419 globular cluster. Admittedly, the terms of such a classification problem are not limited to the considered dataset but, on the contrary, are ubiquitous in astronomy and, more in general, might arise in any field where data components are incommensurable.

Bayesian nonparametric methods for density estimation and clustering have been successfully applied in a wide range of fields, including genetics (Huelsenbeck and Andolfatto, 2007), bioinformatics (Medvedovic and Sivaganesan, 2002), clinical trials (Xu et al., 2017), econometrics (Otranto and Gallo, 2002), to cite but a few. In this work we focus on the Dirichlet process mixture (DPM) model introduced by Lo (1984), arguably the most popular Bayesian nonparametric model. Although its properties have been thoroughly studied (see, e.g., Hjort et al., 2010), little attention has been dedicated to its robustness under data transformations (see Arbel and Nipoti, 2013). To the best of our knowledge, only Bean et al. (2016) study the effect of data transformation under a DPM model: their goal is to transform the sample so to facilitate the estimation of univariate densities on a new scale and thus to improve the performance of the methodology.

In this paper we investigate the effect of affine transformations of the data on location-scale DPM of multivariate Gaussians (DPM-G) (Müller et al., 1996), which will be introduced in Section 2. This is a very commonly used class of DPM models whose asymptotic properties have been studied by Wu and Ghosal (2010), Shen et al. (2013) and Canale and De Blasi (2017), among others. While rescaling the data, often for numerical convenience, is a common practice, the robustness of multivariate DPM-G models under such transformations remains essentially unaddressed to date. We fill this gap by formally studying robustness properties for a flexible specification of DPM-G models, under affine transformation of the data. Specifically, our contribution is two-fold: first, we formalise the intuitive idea that a location-scale DPM-G model on a given dataset induces a location-scale DPM-G model on rescaled data and we provide the parameters mapping for the transformed DPM-G model; second, we introduce the notion of asymptotic robustness under affine transformations of the data and show that, under mild assumptions on the true data generating process, DPM-G models feature such robustness property. Our theoretical results are supported by an extensive simulation study, focusing on both density and clustering estimation. These findings make the DPM-G model a suitable candidate to deal with problems where an informed choice of the relative scale of different dimensions seems prohibitive. We thus fit a DPM-G model to the NGC 2419 dataset and show that it provides interesting insight on the classification problem motivating this work.

The rest of the paper is organised as follows. In Section 2 we describe the modelling framework and introduce the notation used throughout the paper. Sections 3 and 4 present the main results of the work, with respective focus on finite sample properties and large sample asymptotics. A thorough simulation study is presented in Section 5 while Section 6 is dedicated to the analysis of

the NGC 2419 dataset. Conclusions are discussed in Section 7. Finally, proofs of two technical lemmas are postponed to Appendix A.

2. Modelling framework

Let $\mathbf{X}^{(n)} := (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be a sample of size n of d-dimensional observations $\mathbf{X}_i := (X_{i,1}, \ldots, X_{i,d})^\intercal$ defined on some probability space $(\Omega, \mathscr{A}, \mathbb{P})$ and taking values in $\mathbb{R}^d$. Consider an invertible affine transformation $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$, that is $g(\mathbf{x}) = \mathbf{C}\mathbf{x} + \mathbf{b}$ where $\mathbf{C}$ is an invertible matrix of dimension $d \times d$ and $\mathbf{b}$ a d-dimensional column vector. The nature of the transformation g is such that, if applied to a random vector $\mathbf{X}$ with probability density function f, it gives rise to a new random vector $g(\mathbf{X})$ with probability density function $f_g = |\det(\mathbf{C})|^{-1} f \circ g^{-1}$.

Henceforth we denote by $\mathscr{F}$ the space of all density functions with support on $\mathbb{R}^d$. The DPM model (Lo, 1984) defines a random density taking values in $\mathscr{F}$ as

$$\tilde{f}(\mathbf{x}) = \int_\Theta k(\mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\tilde{P}(\boldsymbol{\theta}) \tag{1}$$

where $k(\mathbf{x}; \boldsymbol{\theta})$ is a kernel on $\mathbb{R}^d$ parameterized by $\boldsymbol{\theta} \in \Theta$, $\tilde{P}$ is a Dirichlet process (DP) with parameters α (precision parameter) and $P_0 := \mathbb{E}[\tilde{P}]$ (base measure), a distribution defined on Θ (Ferguson, 1973). The almost sure discreteness of $\tilde{P}$ allows the random density $\tilde{f}$ to be rewritten as

$$\tilde{f}(\mathbf{x}) = \sum_{i=1}^{\infty} w_i k(\mathbf{x}; \boldsymbol{\theta}_i), \tag{2}$$

where the random atoms $\boldsymbol{\theta}_i$ are i.i.d. from P_0, and the random jumps w_i, independent of the atoms, admit the following stick-breaking representation (Sethuraman, 1994): given a set of random weights $v_i \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$ (independent of the atoms $\boldsymbol{\theta}_i$), then $w_1 = v_1$ and, for $j \geq 2$, $w_j = v_j \prod_{i=1}^{j-1}(1 - v_i)$. While several kernels $k(\mathbf{x}; \boldsymbol{\theta})$ have been considered in the literature, including e.g. skew-normal (Canale and Scarpa, 2016), Weibull (Kottas, 2006), Poisson (Krnjajić et al., 2008), here we focus on the convenient and commonly adopted Gaussian specification of Escobar and West (1995) and Müller et al. (1996). In the latter case, $k(\mathbf{x}; \boldsymbol{\theta})$ represents a d-dimensional Gaussian kernel $\phi_d(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, provided that $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the column vector $\boldsymbol{\mu}$ and the matrix $\boldsymbol{\Sigma}$ represent, respectively, mean vector and covariance matrix of the Gaussian kernel. This specification defines the model referred to as d-dimensional location-scale Dirichlet process mixture of Gaussians (DPM-G), which can be represented in hierarchical form as

$$\begin{aligned} \mathbf{X}_i \mid \boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) &\overset{\text{ind}}{\sim} \phi_d(\mathbf{x}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \\ \boldsymbol{\theta}_i \mid \tilde{P} &\overset{\text{iid}}{\sim} \tilde{P}, \\ \tilde{P} &\sim DP(\alpha, P_0). \end{aligned} \tag{3}$$

The almost sure discreteness of $\tilde{P}$ implies that the vector $\boldsymbol{\theta}^{(n)} := (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ might show ties with positive probability, thus leading to a partition of $\boldsymbol{\theta}^{(n)}$ into $K_n \leq n$ distinct values. This, in turn, leads to a partition of the set of observations $\mathbf{X}^{(n)}$, obtained by grouping two observations $\mathbf{X}_{i_1}$ and $\mathbf{X}_{i_2}$ together if and only if $\boldsymbol{\theta}_{i_1} = \boldsymbol{\theta}_{i_2}$. This observation implies that the posterior distribution of the random density $\tilde{f}$ carries useful information on the clustering structure of the data, thus making DPM-G models convenient tools for density and clustering estimation problems.

Although other specifications for the base measure can be considered (see, e.g., Görür and Rasmussen, 2010), we choose to work within the framework set forth by Müller et al. (1996) where P_0 is defined as the product of two independent distributions for the location parameter $\boldsymbol{\mu}$ and the scale parameter $\boldsymbol{\Sigma}$, namely a multivariate normal and an inverse-Wishart distribution, that is

$$P_0(\mathrm{d}\boldsymbol{\mu}, \mathrm{d}\boldsymbol{\Sigma}; \boldsymbol{\pi}) = N_d(\mathrm{d}\boldsymbol{\mu}; \mathbf{m}_0, \mathbf{B}_0) \times IW(\mathrm{d}\boldsymbol{\Sigma}; \nu_0, \mathbf{S}_0). \tag{4}$$

For the sake of compactness, we use the notation $\boldsymbol{\pi} := (\mathbf{m}_0, \mathbf{B}_0, \nu_0, \mathbf{S}_0)$ to denote the vector of hyperparameters characterising the base measure P_0. We denote by Π the prior distribution induced on $\mathscr{F}$ by the DPM-G model (2) with base measure (4).

3. DPM-G model and affine transformation of the data

Let $\tilde{f}_{\boldsymbol{\pi}}$ be a DPM-G model defined as in (2), with base measure (4) and hyperparameters $\boldsymbol{\pi}$. The next result shows that, for any invertible affine transformation $g(\mathbf{x}) = \mathbf{C}\mathbf{x} + \mathbf{b}$, there exists a specification $\boldsymbol{\pi}_g := (\mathbf{m}_0^{(g)}, \mathbf{B}_0^{(g)}, \nu_0^{(g)}, \mathbf{S}_0^{(g)})$ of the hyperparameters characterising the base measure in (4), such that $\tilde{f}_{\boldsymbol{\pi}_g} = |\det(\mathbf{C})|^{-1} \tilde{f}_{\boldsymbol{\pi}} \circ g^{-1}$. That is, for every $\omega \in \Omega$ and given a random vector $\mathbf{X}$ distributed according to $\tilde{f}_{\boldsymbol{\pi}}(\omega)$, we have that $\tilde{f}_{\boldsymbol{\pi}_g}(\omega)$ is the density of the transformed random vector $g(\mathbf{X})$.

Proposition 1. *Let $\tilde{f}_{\boldsymbol{\pi}}$ be a location-scale DPM-G model defined as in (2), with base measure (4) and hyperparameters $\boldsymbol{\pi} = (\mathbf{m}_0, \mathbf{B}_0, \nu_0, \mathbf{S}_0)$. For any invertible affine transformation $g(\mathbf{x}) = \mathbf{C}\mathbf{x} + \mathbf{b}$, we have*

$$\tilde{f}_{\boldsymbol{\pi}_g} = |\det(\mathbf{C})|^{-1} \tilde{f}_{\boldsymbol{\pi}} \circ g^{-1},$$

where $\boldsymbol{\pi}_g := (\mathbf{C}\mathbf{m}_0 + \mathbf{b}, \mathbf{C}\mathbf{B}_0\mathbf{C}^\intercal, \nu_0, \mathbf{C}\mathbf{S}_0\mathbf{C}^\intercal)$.

Proof. Model $\tilde{f}_{\boldsymbol{\pi}}$ can be written as

$$\begin{aligned}
\tilde{f}_{\boldsymbol{\pi}}(\mathbf{x}) &= \int (2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tilde{P}(\mathrm{d}\boldsymbol{\mu}, \mathrm{d}\boldsymbol{\Sigma}; \boldsymbol{\pi}) \\
&= \int (2\pi)^{-\frac{d}{2}} |\det(\mathbf{C})| \det(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\intercal)^{-\frac{1}{2}} \\
&\quad \times \exp\left\{-\frac{1}{2}(\mathbf{C}\mathbf{x} + \mathbf{b} - \mathbf{C}\boldsymbol{\mu} - \mathbf{b})^\intercal (\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\intercal)^{-1} (\mathbf{C}\mathbf{x} + \mathbf{b} - \mathbf{C}\boldsymbol{\mu} - \mathbf{b})\right\} \tilde{P}(\mathrm{d}\boldsymbol{\mu}, \mathrm{d}\boldsymbol{\Sigma}; \boldsymbol{\pi}).
\end{aligned}$$

By performing the change of variables $\mathbf{S} = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\intercal$ and $\mathbf{m} = \mathbf{C}\boldsymbol{\mu} + \mathbf{b}$ and observing that, by standard properties of the inverse-Wishart and normal distributions,

1. $\boldsymbol{\Sigma} \sim IW(\nu_0, \mathbf{S}_0)$ implies $\mathbf{S} \sim IW(\nu_0, \mathbf{C}\mathbf{S}_0\mathbf{C}^\intercal)$,
2. $\boldsymbol{\mu} \sim N_d(\mathbf{m}_0, \mathbf{B}_0)$ implies $\mathbf{m} \sim N_d(\mathbf{C}\mathbf{m}_0 + \mathbf{b}, \mathbf{C}\mathbf{B}_0\mathbf{C}^\intercal)$,
3. $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ implies $\mathbf{C}\mathbf{X} + \mathbf{b} \sim N_d(\mathbf{m}, \mathbf{S})$,

we obtain

$$\begin{aligned}\tilde{f}_{\boldsymbol{\pi}}(\mathbf{x}) &= |\det(\mathbf{C})| \int (2\pi)^{-\frac{d}{2}} \det(\mathbf{S})^{-\frac{1}{2}} \\ &\quad \times \exp\left\{-\frac{1}{2}(\mathbf{C}\mathbf{x} + \mathbf{b} - \mathbf{m})^\intercal \mathbf{S}^{-1}(\mathbf{C}\mathbf{x} + \mathbf{b} - \mathbf{m})\right\} \tilde{P}(\mathrm{d}\mathbf{m}, \mathrm{d}\mathbf{S}; \boldsymbol{\pi}_g) \\ &= |\det(\mathbf{C})| \tilde{f}_{\boldsymbol{\pi}_g}(g(\mathbf{x})).\end{aligned}$$

A simple reparametrisation leads to $\tilde{f}_{\boldsymbol{\pi}_g} = |\det(\mathbf{C})|^{-1} \tilde{f}_{\boldsymbol{\pi}} \circ g^{-1}$. All the identities in this proof are deterministic, that is they hold for every $\omega \in \Omega$. □

This result implies that, for any invertible affine transformation g, modelling the set of observations $\mathbf{X}^{(n)}$ with a DPM-G model (2), with base measure (4) and hyperparameters $\boldsymbol{\pi}$, is equivalent with assuming the same model with transformed hyperparameters $\boldsymbol{\pi}_g$, for the transformed observations $g(\mathbf{X})^{(n)} := (g(\mathbf{X}_1), \ldots, g(\mathbf{X}_n))$. As a by-product, the same posterior inference can be drawn conditionally on both the original and the transformed set of observations, as the conditional distribution of the random density $\tilde{f}_{\boldsymbol{\pi}_g}$, given $g(\mathbf{X})^{(n)}$, coincides with the conditional distribution of $|\det(\mathbf{C})|^{-1} \tilde{f}_{\boldsymbol{\pi}} \circ g^{-1}$, given $\mathbf{X}^{(n)}$. Proposition 1 thus provides a formal justification for the procedure of transforming data, e.g. via standardisation or normalisation, often adopted to achieve numerical efficiency: as long as the prior specification of the hyperparameters of a DPM-G model respects the condition of Proposition 1, transforming the data does not affect posterior inference.

The elicitation of an honest prior, thus independent of the data, for the hyperparameters $\boldsymbol{\pi}$ of the base measure (4) of a DPM model is in general a difficult task. A popular practice, therefore, consists in setting the hyperparameters equal to some empirical estimates $\hat{\boldsymbol{\pi}}(\mathbf{X}^{(n)})$, by applying the so-called empirical Bayes approach (see, e.g., Lehmann and Casella, 2006). Recent investigations (Petrone et al., 2014; Donnet et al., 2018) provide a theoretical justification of this hybrid procedure by shedding light on its asymptotic properties. The next example shows that this procedure satisfies the assumptions of Proposition 1 and, thus, guarantees that posterior Bayesian inference, under an empirical Bayes approach, is not affected by affine transformations to the data.

Example 1 (Empirical Bayes approach)**.** A commonly used empirical Bayes approach for specifying the hyperparameters $\boldsymbol{\pi}$ of a DPM-G model, defined as in (2) and (4), consists in setting

$$\mathbf{m}_0 = \overline{\mathbf{X}}, \qquad \mathbf{B}_0 = \frac{1}{\gamma_1}\mathbf{S}^2_{\mathbf{X}}, \qquad \mathbf{S}_0 = \frac{\nu_0 - d - 1}{\gamma_2}\mathbf{S}^2_{\mathbf{X}}, \tag{5}$$

where $\overline{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i/n$ and $\mathbf{S}^2_{\mathbf{X}} = \sum_{i=1}^n (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^\intercal/(n-1)$ are the sample mean vector and the sample covariance matrix, respectively, and $\gamma_1, \gamma_2 > 0$, $\nu_0 > d + 1$. This specification for the hyperparameters $\boldsymbol{\pi}$ has a straightforward interpretation. Namely, the parameter $\mathbf{m}_0$, mean of the prior guess distribution of $\boldsymbol{\mu}$, can be interpreted as the overall mean value and, in absence of available prior information, set equal to the observed sample mean. Similarly, the parameter $\mathbf{B}_0$, covariance matrix of the prior guess distribution of $\boldsymbol{\mu}$, is set equal to a penalised version of the sample covariance matrix $\mathbf{S}^2_{\mathbf{X}}$, where γ_1 takes on the interpretation of the size of the ideal prior sample upon which the prior guess on the distribution of $\boldsymbol{\mu}$ is based. Similarly, the hyperparameter $\mathbf{S}_0$ is set equal to a penalised version of the sample covariance matrix $\mathbf{S}^2_{\mathbf{X}}$, choice that corresponds to the prior guess that the covariance matrix of each component of the mixture coincides with a rescaled version of the sample covariance matrix. Specifically, $\mathbf{S}_0 = \mathbf{S}^2_{\mathbf{X}}(\nu_0 - d - 1)/\gamma_2$ follows by setting $\mathbb{E}[\boldsymbol{\Sigma}] = \mathbf{S}^2_{\mathbf{X}}/\gamma_2$ and observing that, by standard properties of the inverse-Wishart distribution, $\mathbb{E}[\boldsymbol{\Sigma}] = \mathbf{S}_0/(\nu_0 - d - 1)$. Finally the parameter ν_0 takes on the interpretation of the size of an ideal prior sample upon which the prior guess $\mathbf{S}_0$ is based. Next we focus on the setting of the hyperparameters $\boldsymbol{\pi}_g$, given the transformed observations $g(\mathbf{X})^{(n)}$. The same empirical Bayes procedure adopted in (5) leads to

$$\mathbf{m}_0^{(g)} = \overline{g(\mathbf{X})} = \mathbf{C}\mathbf{m}_0 + \mathbf{b}, \qquad \mathbf{B}_0^{(g)} = \frac{1}{\gamma_1}\mathbf{S}^2_{g(\mathbf{X})}, \qquad \mathbf{S}_0^{(g)} = \frac{\nu_0 - d - 1}{\gamma_2}\mathbf{S}^2_{g(\mathbf{X})}.$$

Observing that $\mathbf{S}^2_{g(\mathbf{X})} = \mathbf{C}\mathbf{S}^2_{\mathbf{X}}\mathbf{C}^\intercal$ and setting $\nu_0^{(g)} = \nu_0$ shows that the described empirical Bayes procedure corresponds to $\boldsymbol{\pi}_g = (\mathbf{C}\mathbf{m}_0 + \mathbf{b}, \mathbf{C}\mathbf{B}_0\mathbf{C}^\intercal, \nu_0, \mathbf{C}\mathbf{S}_0\mathbf{C}^\intercal)$ and, thus, by Proposition 1, $\tilde{f}_{\boldsymbol{\pi}_g} = |\det(\mathbf{C})|^{-1}\tilde{f}_{\boldsymbol{\pi}} \circ g^{-1}$. □

4. Large n asymptotic robustness

We investigate the effect of affine transformations of the data on DPM-G models by studying the asymptotic behaviour of the resulting posterior distribution in the large sample size regime. To this end, we consider a scenario that mimics a situation where no precise information about the scale of the data is available, and thus the prior model must be specified arbitrarily. More specifically, we fit the same DPM-G model $\tilde{f}_{\boldsymbol{\pi}}$, defined in (2) and (4), to two versions of the data, that is $\mathbf{X}^{(n)}$ and $g(\mathbf{X})^{(n)}$, by using the exact same specification for the hyperparameters $\boldsymbol{\pi}$. Under this setting, the assumptions of Proposition 1 are not met and the posterior distributions obtained by conditioning on the two sets of observations are different random distributions which, thus, might lead to different statistical conclusions. The main result of this section shows that, under

mild conditions on the true generating distribution of the observations, the posterior distributions obtained by conditioning $\tilde{f}_{\boldsymbol{\pi}}$ on the two sets of observations $\mathbf{X}^{(n)}$ and $g(\mathbf{X})^{(n)}$, become more and more similar, up to an affine reparametrisation, as the sample size n grows. More specifically we show that the probability mass of the joint distribution of these two conditional random densities concentrates in a neighbourhood of $\{(f_1, f_2) \in \mathcal{F} \times \mathcal{F} \text{ s.t. } f_1 = |\det(\mathbf{C})| f_2 \circ g\}$ as n goes to infinity. Henceforth we will say that the DPM-G model (2) with base measure (4) is asymptotically robust to affine transformation of the data. We first formalise this result and then provide its proof in Section 4.1. The latter is presented as split into intermediary lemmas whose proofs are deferred to Appendix A.

Henceforth we consider a metric ρ on $\mathscr{F}$ which can be equivalently defined as the Hellinger distance $\rho(f_1, f_2) = \{\int (\sqrt{f_1(\mathbf{x})} - \sqrt{f_2(\mathbf{x})})^2 \mathrm{d}\mathbf{x}\}^{1/2}$ or the L^1 distance $\rho(f_1, f_2) = \int |f_1(\mathbf{x}) - f_2(\mathbf{x}))| \mathrm{d}\mathbf{x}$ between densities f_1 and f_2 in $\mathscr{F}$, and we denote by $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^d$. Moreover, we adopt here the usual frequentist validation approach in the large n regime, working 'as if' the observations $\mathbf{X}^{(n)}$ were generated from a true and fixed data generating process (see for instance Rousseau, 2016). We also assume that this data generating process admits a density function with respect to the Lebesgue measure, denoted by f^*. In the setting we consider, the same model $\tilde{f}_{\boldsymbol{\pi}}$ defined in (2) and (4) is fitted to $\mathbf{X}^{(n)}$ and $g(\mathbf{X})^{(n)}$, thus leading to two distinct posterior random densities, with distributions on $\mathscr{F}$ denoted by $\Pi(\cdot \mid \mathbf{X}^{(n)})$ and $\Pi(\cdot \mid g(\mathbf{X})^{(n)})$, respectively. We use the notation $\Pi_2(\cdot \mid \mathbf{X}^{(n)})$ to refer to their joint posterior distribution on $\mathscr{F} \times \mathscr{F}$.

Theorem 1. Let $f^* \in \mathscr{F}$, true generating density of $\mathbf{X}^{(n)}$, satisfy the conditions

A1. $0 < f^*(\mathbf{x}) < M$, for some constant M and for all $\mathbf{x} \in \mathbb{R}^d$,

A2. $\left|\int f^*(\mathbf{x}) \log f^*(\mathbf{x}) \mathrm{d}\mathbf{x}\right| < \infty$,

A3. $\exists\, \delta > 0$ such that $\int f^*(\mathbf{x}) \log\left(f^*(\mathbf{x})/\varphi_\delta(\mathbf{x})\right) \mathrm{d}\mathbf{x} < \infty$, where $\varphi_\delta(\mathbf{x}) = \inf_{\{\mathbf{t} \,:\, \|\mathbf{t}-\mathbf{x}\| < \delta\}} f^*(\mathbf{t})$,

A4. for some $\eta > 0$, $\int \|\mathbf{x}\|^{2(1+\eta)} f^*(\mathbf{x}) \mathrm{d}\mathbf{x} < \infty$.

Let $g : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ be an invertible affine transformation and $\Pi_2(\cdot \mid \mathbf{X}^{(n)})$ be the joint posterior distribution induced by a DPM-G as (2) with base measure (4) where $\nu_0 > (d+1)(2d-3)$. Then, for any $\varepsilon > 0$,

$$\Pi_2((f_1, f_2) : \rho(f_1, |\det(\mathbf{C})| f_2 \circ g) < \varepsilon \mid \mathbf{X}^{(n)}) \longrightarrow 1$$

as $n \to \infty$.

The assumptions of Theorem 1 thus refer to the true generating distribution f^* of $\mathbf{X}^{(n)}$. Assumptions A1 and A2 require f^* to be bounded, fully supported on $\mathbb{R}^d$ and with finite entropy. Note that Assumption A2 is not implied by Assumption A1: for instance, if the true density $f^*(\mathbf{x})$ is defined on $\mathbb{R}$ with a right tail behaving as $\mathbf{x}^{-1}(\log \mathbf{x})^{-2}$ at infinity, then $f^*(\mathbf{x}) \log f^*(\mathbf{x})$ behaves like

$(\mathbf{x}\log\mathbf{x})^{-1}$, and the entropy is infinite.[1] Assumption A3 is a condition of local regularity of the entropy of f^*. Finally, Assumption A4 requires the tails of f^* to be thin enough for some moment of order strictly larger than two to exist.

4.1. Proof of Theorem 1

The proof relies on results proved by Canale and De Blasi (2017). Let $\boldsymbol{\lambda}(\boldsymbol{\Sigma}^{-1}) := (\lambda_1(\boldsymbol{\Sigma}^{-1}), \ldots, \lambda_d(\boldsymbol{\Sigma}^{-1}))$ be the vector of eigenvalues, in increasing order, of $\boldsymbol{\Sigma}^{-1}$, the precision matrix of the Gaussian kernel. Henceforth we write $f(x) \lesssim g(x)$ to indicate that the inequality $f(x) \leq cg(x)$ holds for some constant c and for any x.

Theorem 2. (Theorem 2 in Canale and De Blasi, 2017). Let $f^* \in \mathscr{F}$, true generating density of $\mathbf{X}^{(n)}$, satisfy the conditions of Theorem 1, and model $\mathbf{X}^{(n)}$ by means of a DPM-G model defined in (2). Suppose that the base measure P_0 has the product form $P_0(\mathrm{d}\boldsymbol{\mu}, \mathrm{d}\boldsymbol{\Sigma}) = P_{0,1}(\mathrm{d}\boldsymbol{\mu})P_{0,2}(\mathrm{d}\boldsymbol{\Sigma})$ and that $P_{0,1}$ and $P_{0,2}$ satisfy the following conditions: for some positive constants c_1, c_2, c_3, $r > (d-1)/2$ and $\kappa > d(d-1)$,

B1. $P_{0,1}(\|\boldsymbol{\mu}\| > x) \lesssim x^{-2(r+1)}$,

B2. $P_{0,2}(\lambda_d(\boldsymbol{\Sigma}^{-1}) > x) \lesssim \exp\{-c_1 x^{c_2}\}$,

B3. $P_{0,2}\left(\lambda_1(\boldsymbol{\Sigma}^{-1}) < \frac{1}{x}\right) \lesssim x^{-c_3}$,

B4. $P_{0,2}\left(\frac{\lambda_d(\boldsymbol{\Sigma}^{-1})}{\lambda_1(\boldsymbol{\Sigma}^{-1})} > x\right) \lesssim x^{-\kappa}$,

all for any sufficiently large x. Then the posterior distribution $\Pi(\cdot|\mathbf{X}^{(n)})$ is consistent at f^*, that is, for every $\varepsilon > 0$,

$$\Pi\left(f : \rho(f, f^*) < \varepsilon \mid \mathbf{X}^{(n)}\right) \longrightarrow 1$$

as $n \to \infty$.

Theorem 2 provides general conditions on the base measure P_0 which guarantee consistency of the posterior distribution. The next lemma shows that these conditions are met by the normal/inverse-Wishart base measure (4).

Lemma 1. *Conditions B1–B4 of Theorem 2 are satisfied by the multivariate normal/inverse-Wishart base measure (4) with $\nu_0 > (d+1)(2d-3)$.*

Although the proof of Lemma 1 can be found in Canale and De Blasi (2017) (Corollary 1, relying, in turn, on results by Shen et al. (2013)), we provide it in Appendix A for the sake of completeness and in order to account for the slightly different prior specification considered in this work. Next lemma shows that if f^* satisfies conditions A1–A4 of Theorem 1, so does $f_g^* := |\det(\mathbf{C})|^{-1} f^* \circ g^{-1}$, for any invertible affine transformation g.

[1] The function $\mathbf{x} \mapsto \mathbf{x}^{-a}(\log\mathbf{x})^{-b}$ is integrable at infinity if and only if $a > 1$ or $a = 1$ and $b > 1$.

Lemma 2. *If conditions A1–A4 of Theorem 1 are satisfied by f^*, then for any invertible affine transformation $g(\mathbf{x}) = \mathbf{C}\mathbf{x} + \mathbf{b}$, they are also satisfied by f_g^*.*

The proof of Leamma 2 is postponed to Appendix A. We are now ready to prove Theorem 1 by combining Theorem 2 with Lemma 1 and Lemma 2.

Proof of Theorem 1. By combining Lemma 1, Lemma 2 and Theorem 2, we have that for any $\epsilon > 0$,

$$\Pi\left(f : \rho(f, f^*) < \epsilon/2 \mid \mathbf{X}^{(n)}\right) \longrightarrow 1, \tag{6}$$

$$\Pi\left(f : \rho(f, f_g^*) < \epsilon/2 \mid g(\mathbf{X})^{(n)}\right) \longrightarrow 1, \tag{7}$$

as $n \to \infty$. We notice that the distance ρ is invariant with respect to change of variables and thus $\rho(|\det(\mathbf{C})|^{-1} f_2 \circ g^{-1}, f^*) = \rho(f_2, f_g^*)$. This, combined with the triangular inequality, leads to

$$\begin{aligned}
&\Pi_2((f_1, f_2) : \rho(f_1, |\det(\mathbf{C})|^{-1} f_2 \circ g^{-1}) < \epsilon \mid \mathbf{X}^{(n)}) \\
&\geq \Pi_2\left((f_1, f_2) : \rho(f_1, f^*) < \epsilon/2,\ \rho(f_2, f_g^*) < \epsilon/2 \mid \mathbf{X}^{(n)}\right) \\
&\geq \Pi_2\left((f_1, f_2) : \rho(f_1, f^*) < \epsilon/2 \mid \mathbf{X}^{(n)}\right) + \Pi_2\left((f_1, f_2) : \rho(f_2, f_g^*) < \epsilon/2 \mid \mathbf{X}^{(n)}\right) - 1 \\
&= \Pi\left(f_1 : \rho(f_1, f^*) < \epsilon/2 \mid \mathbf{X}^{(n)}\right) + \Pi\left(f_2 : \rho(f_2, f_g^*) < \epsilon/2 \mid g(\mathbf{X})^{(n)}\right) - 1 \\
&\quad \longrightarrow 1 + 1 - 1 = 1,
\end{aligned}$$

as $n \to \infty$. As a result, for $n \to \infty$,

$$\Pi_2((f_1, f_2) : \rho(f_1, |\det(\mathbf{C})| f_2 \circ g) < \varepsilon \mid \mathbf{X}^{(n)}) \longrightarrow 1.$$

□

5. Simulation study

We performed a simulation study to provide empirical support to our results on the large n asymptotic robustness of a DPM-G model specified as in (2) with base measure (4), under affine transformations of the data. We considered 15 different simulation scenarios. Specifically, we considered three different sample sizes, namely $n = 100$, $n = 300$ and $n = 1\,000$. Then, for each sample size, we generated a sample from a mixture of two Gaussian components, one being highly correlated and the other uncorrelated, defined as

$$\mathbf{X}^{(n)} \sim \frac{1}{2} N_2\left(\begin{bmatrix}-2\\-2\end{bmatrix}, \begin{bmatrix}1 & 0.85\\0.85 & 1\end{bmatrix}\right) + \frac{1}{2} N_2\left(\begin{bmatrix}2\\2\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right). \tag{8}$$

In order to test the robustness of the model under affine transformations of the data, we stretched or compressed the generated datasets by using five different constants, namely $c = 1/5$, $c = 1/2$, $c = 1$, $c = 2$ and $c = 5$.

For each constant, we multiplied the simulated data by c, thus obtaining a transformed dataset $\mathbf{X}_c^{(n)} := c\mathbf{X}^{(n)}$. For each simulation scenario, namely $c \in \{1/5, 1/2, 1, 2, 5\}$, $n \in \{100, 300, 1\,000\}$, we generated 100 replicates. We then fitted a DPM-G model, specified as in (2) and (4), to each one of the 1 500 simulated datasets. In order to enhance the flexibility of the model, we completed its specification by setting a normal/inverse-Wishart prior distribution for the hyperparameters $(\mathbf{m}_0, \mathbf{B}_0)$ of the base measure (4). Namely, we set $\mathbf{B}_0 \sim IW(4, \text{diag}(\mathbf{15}))$ and $\mathbf{m}_0 \mid \mathbf{B}_0 \sim N(0, \mathbf{B}_0)$, specification chosen so that $\mathbb{E}[\boldsymbol{\mu}] = \mathbf{0}$ and to guarantee a prior guess on the location component $\boldsymbol{\mu}$ flat enough to cover the support of the non-transformed data. As for the scale component of the base measure (4), we set $(\nu_0, \mathbf{S}_0) = (4, \text{diag}(\mathbf{1}))$. Finally, the mass parameter α of the Dirichlet process was set equal to 1.

Realisations of the mean of the posterior distribution were obtained by means of a Gibbs sampler relying on a Blackwell–McQueen Pólya urn scheme (see Müller et al., 1996), implemented in the BNPmix R package[2]. For each replicate, posterior inference was drawn based on 5 000 iterations, obtained after discarding the first 2 500. Convergence of the chains was assessed by visually investigating traceplots referring to randomly selected replicates, which did not provide indication against it.

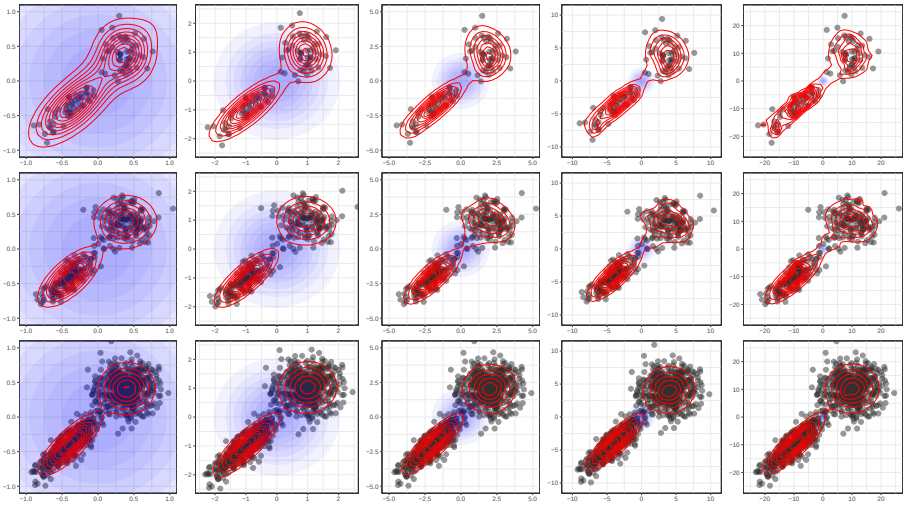

Figure 2: Simulation study. Based on a single replicate of the samples $\mathbf{X}^{(100)}$, $\mathbf{X}^{(300)}$ and $\mathbf{X}^{(1000)}$, scatter plot of the data (grey dots), contour plot of the estimated densities based on a DPM-G model (red curves) and contour plot for the expected prior density (blue filled curves). Left to right: rescaling constant $c = 1/5$, $c = 1/2$, $c = 1$, $c = 2$, $c = 5$. Top to bottom: sample size $n = 100$, $n = 300$, $n = 1000$.

[2]The package is available at https://github.com/rcorradin/BNPmix and can be installed via devtools. For reproducibility, the code is available at https://github.com/rcorradin/Affine.

Figure 2 shows, for every $n \in \{100, 300, 1000\}$ and $c \in \{1/5, 1/2, 1, 2, 5\}$, a contour plot of the estimated posterior densities. The difference between estimated densities, across different values of c, is apparent when $n = 100$, with the two extreme cases, namely $c = 1/5$ and $c = 5$, suggesting a different number of modes in the estimated density. For larger sample sizes, this difference is less evident and, when $n = 1\,000$, the contour plots are hardly distinguishable. These qualitative observations are in agreement with the large n asymptotic results of Theorem 1. The plots of Figure 2 refer to a single realisation of the samples $\mathbf{X}^{(100)}$, $\mathbf{X}^{(300)}$ and $\mathbf{X}^{(100)}$ considered in the simulation study, although qualitatively similar results can be found in almost any replicate.

The findings drawn from a visual inspection of Figure 2 were confirmed by assessing the distance between estimated posterior densities. Specifically, for any considered sample size n and for any pair of values c_1 and c_2 taken by the constant c, we approximately evaluated the L^1 distance between the suitably rescaled estimated posterior densities obtained conditionally on $\mathbf{X}_{c_1}^{(n)}$ and on $\mathbf{X}_{c_2}^{(n)}$. The results of such analysis are shown in Figure 3 and indicate that as the sample size grows, the difference in terms of L^1 distance strictly decreases.

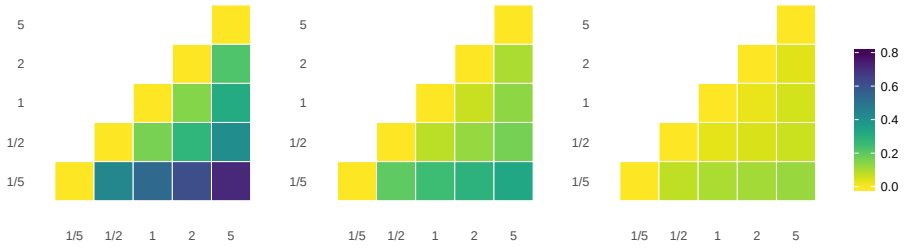

Figure 3: Simulation study. L^1 distance between suitably rescaled estimated densities after data transformations for different constants c_1 (X axis) and c_2 (Y axis), averaged over 100 replications. Left to right: sample size $n = 100$, sample size $n = 300$, sample size $n = 1000$.

The posterior distribution of the random density induced by a DPM-G model provides interesting insight also on the clustering structure of the data. The second goal of the simulation study, thus, consisted in investigating the impact of the scaling factor c on the estimated number of groups in the partition induced on the data. To this end, for each considered n and c, we estimated $\hat{K}_n^{(\mathrm{VI})}$, the number of groups in the optimal partition estimated using a procedure introduced by Wade and Ghahramani (2018) and based on the variation of information loss function. The average values for this quantity, over 100 replicates, are reported in Table 1. There appears to be a clear trend suggesting that a larger scaling constant c leads to a larger $\hat{K}_n^{(\mathrm{VI})}$: this finding is consistent with the fact that, if the data are stretched while the prior specification is kept unchanged, then we expect the estimated posterior density to need a larger number of Gaussian components to cover the support of the sample. For the purpose of this simulation study the main quantity of interest is the ratio

	$c = 1/5$	$c = 1/2$	$c = 1$	$c = 2$	$c = 5$
$n = 100$	1.81	2.04	2.84	5.96	10.52
$n = 300$	2.00	2.03	2.20	2.82	5.18
$n = 1000$	2.00	2.00	2.04	2.05	2.12

Table 1: Simulation study. Averages over 100 replicates for $\hat{K}_n^{(\mathrm{VI})}$, the number of clusters of the estimated partition estimated by means of Wade and Ghahramani (2018)'s variation of information method. Left to right: rescaling constant $c = 1/5$, $c = 1/2$, $c = 1$, $c = 2$, $c = 5$. Top to bottom: sample size $n = 100$, $n = 300$, $n = 1\,000$.

between the estimated number of groups under any two distinct values c_1 and c_2 for the scaling constant c, that is $\hat{K}_{n,c_1}^{(\mathrm{VI})} / \hat{K}_{n,c_2}^{(\mathrm{VI})}$. The results presented in Table 1 clearly indicate that, as the sample size n becomes large, such ratios tend to approach 1. This suggests that the large n robustness property of the DPM-G model nicely translates to an equivalent notion of robustness in terms of the estimated number of groups $\hat{K}_n^{(\mathrm{VI})}$ in the data.

6. Astronomical data

The large n asymptotic robustness to affine transformation of the DPM-G model makes it a suitable candidate also for analysing data whose components are not commensurable and for which an informed choice of the relative scale of different dimensions seems prohibitive. We fitted the DPM-G model, specified as in (2) and with base measure (4), to the NGC 2419 dataset described in Section 1. The ultimate goal of our analysis consists in classifying stars as belonging to the NGC 2419 globular cluster or as being contaminants: an accurate classification is crucial for the astronomers to study the dynamics of the globular cluster. Since the large majority of the stars in the dataset is expected to belong to the globular cluster, with only a few of them being contaminants, we will identify the globular cluster as the largest group in the estimated partition of the dataset.

Prior to any analysis, data were standardised component by component, the legitimacy of such procedure following from the robustness results of Theorem 1. Hyperprior distributions were specified for the location parameter of the base measure (4) and on the DP mass parameter α. Specifically, $\mathbf{B}_0 \sim IW(6, \mathrm{diag}(\mathbf{15}))$ and $\mathbf{m}_0 \mid \mathbf{B}_0 \sim N(0, \mathbf{B}_0)$, specification chosen to guarantee a prior guess on the location component $\boldsymbol{\mu}$ flat enough to cover the support of the data and centered at $\mathbf{0}$. In addition, the precision parameter α was given a gamma prior distribution with unit shape parameter and rate parameter equal to 5.26, so to reflect the prior opinion of astronomers who would expect two distinct groups of stars in the dataset. Finally, as far as the scale component of the base measure (4) is concerned, we set $(\nu_0, \mathbf{S}_0) = (26, \mathrm{diag}(\mathbf{21}))$, where the number of degrees of freedom $\nu_0 = 26$ of the inverse-Wishart distribution was chosen to guarantee the conditions of Theorem 1 and, in turn, the scale matrix $\mathbf{S}_0 = \mathrm{diag}(\mathbf{21})$ so that $\mathbb{E}[\boldsymbol{\Sigma}] = \mathrm{diag}(\mathbf{1})$. Realisations of the mean of the posterior

distribution were obtained by means of a Gibbs sampler relying on a Blackwell–McQueen Pólya urn scheme[3]. In turn, posterior inference was drawn based on 20 000 iterations, after a burn-in period of 5 000 iterations. Convergence of the chains was assessed by visually investigating traceplots, which did not provide indication against it.

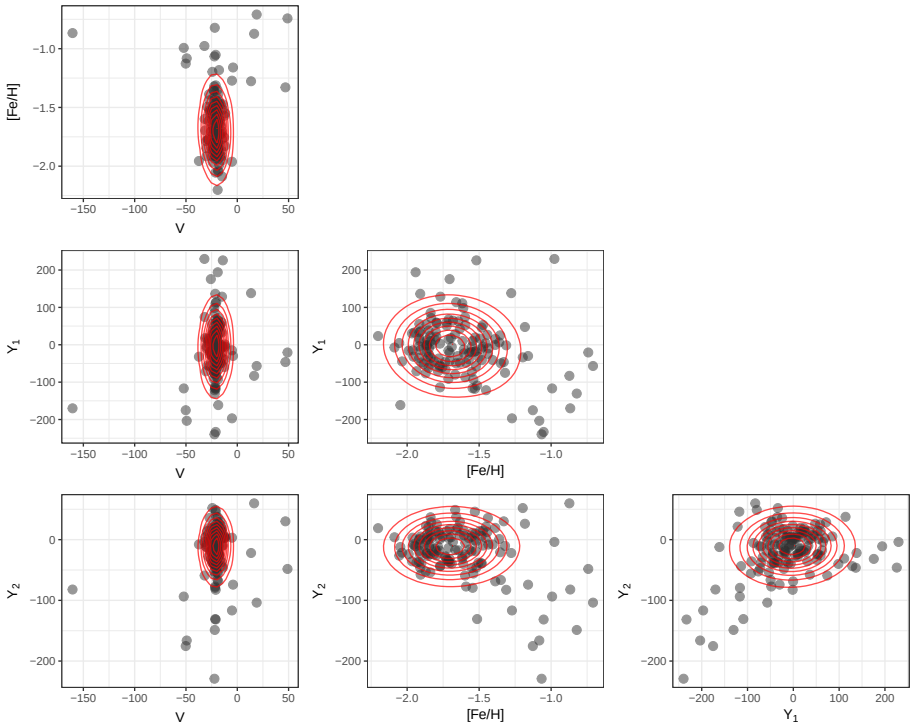

Figure 4: NGC 2419 data. Contour plots of the bivariate marginal densities estimated via DPM-G model.

Figure 4 displays contour plots for the six two-dimensional projections of the estimated posterior density, while Figure 5 shows the scatter plots of the dataset with individual observations coloured according to their membership in the partition estimated based on the variation of information loss function (Wade and Ghahramani, 2018) and labeled as main group (grey circles) and other groups (coloured triangles). The estimated partition is composed of five groups. The largest one, identified as the globular cluster, consists of 124 stars. The remaining 15 stars are thus considered contaminants and are further divided into four groups, one composed by eight stars (group A), one containing five stars (group B) and two singletons (groups C and D). A visual investigation of Figure 5 suggests that stars in group A differ from those in the globular cluster

[3]See footnote 2.

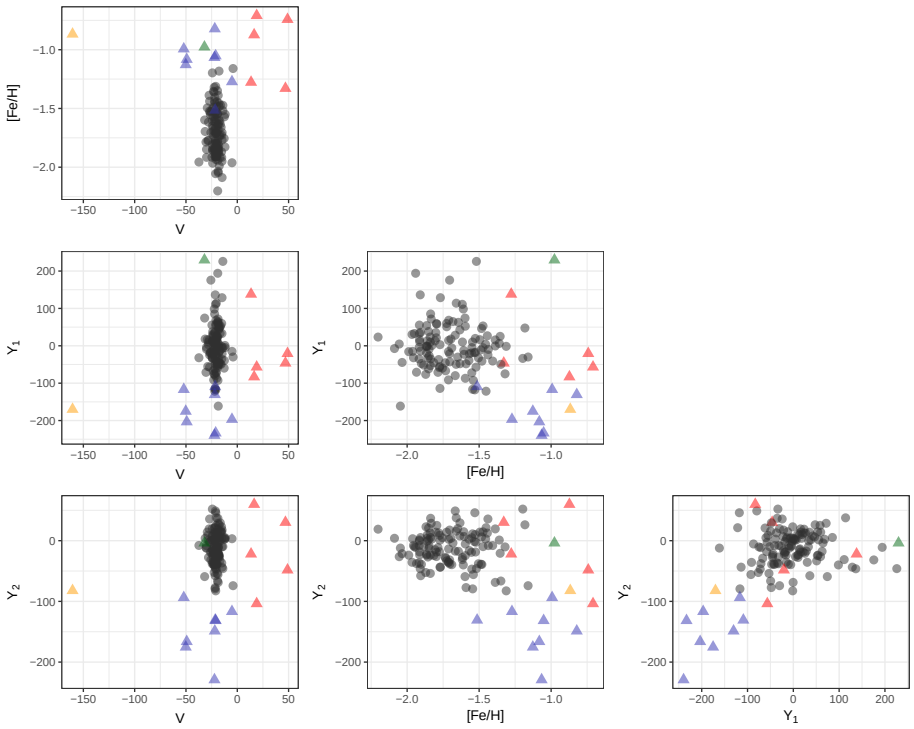

Figure 5: NGC 2419 data. Partition estimated via DPM-G models combined with Wade and Ghahramani (2018)'s variation of information method. Five groups are detected: the largest group (grey dots), group A (blue triangles), group B (red triangles), group C (one orange triangle), group D (one green triangle).

in terms of metallicity and position, with the contaminants characterised by larger values for [Fe/H] and smaller values for Y_1 and Y_2. The stars in group B differ from the globular cluster in terms of velocity and metallicity, with the contaminants showing larger values for V and [Fe/H]). Finally, groups C and D are singletons, the first one being characterised by a high metallicity and an extremely small value for the velocity, the second one showing large values for both metallicity and location Y_1.

Our unsupervised statistical clustering can be compared to the clustering of Ibata et al. (2011) (described in their Figure 4) based on ad hoc physical considerations. Specifically, once the best fitting physical model, in the class of either Newtonian or Modified Newtonian Dynamics models, is detected, they use it in order to compute the average values of the physical variables describing the stars. Stars are then assigned to the globular cluster based on a comparison between their velocity and the average model velocity: those lying close enough are deemed to belong to the cluster, while the others are considered as potential contaminants. For the latter, the evidence of being contaminants is measured by evaluating how distant their metallicity is from the average model one. Two

classifications are then proposed: the first one assigns to the globular cluster only the 118 stars for which the evidence seems strong, the second and less conservative strategy classifies as belonging to the globular cluster a total of 130 stars. Following this distinction and for the sake of simplicity, we summarise the results of Ibata et al. (2011)'s analysis, by devising three groups of stars:

- *globular cluster*: 118 stars deemed to belong to the globular cluster,
- *likely globular cluster*: 12 stars assigned to the globular cluster only when the less conservative procedure is adopted,
- *contaminants*: 9 stars with strong evidence of being contaminants.

			DPM-G groups				
			largest	*A*	*B*	*C*	*D*
		total	*124*	*8*	*5*	*1*	*1*
Ibata et al. groups	*globular cluster*	*118*	114	4	0	0	0
	likely globular cluster	*12*	10	1	0	0	1
	contaminants	*9*	0	3	5	1	0

Table 2: NGC 2419 data. Comparison between the groups identified by Ibata et al. (2011) and the groups estimated via DPM-G model.

For the purpose of comparison, we report in Table 2 the confusion matrix of the groups obtained via the DPG-G model against the groups detected by Ibata et al. All of the 124 stars belonging to the largest group of the partition estimated based on the DPM-G model belong to the groups identified as *globular cluster* or *likely globular cluster* by Ibata et al. At the same time, out of the nine stars classified as contaminants by Ibata et al., the approach based on the DPM-G model assigns none to the globular cluster, three to group A, five stars to group B, which is composed only by stars considered contaminants in Ibata et al., and the star of group C, which shows an extremely small value for the velocity variable. Finally, the group D contains only one star, which is not consider a contaminant in Ibata et al.

Further insight on the clustering structure of the data is provided by Figure 6, which shows the heatmap representation of the posterior similarity matrix obtained from the MCMC output. In agreement with the partition obtained by applying the approach of Wade and Ghahramani (2018), one main group identified with the globular cluster can be clearly detected in Figure 6. As for the remaining stars, arguably the contaminants, there seems to be two well defined groups, A and B, and a few stars whose group membership is less certain.

7. Conclusion

The purpose of this paper was to investigate the behaviour of the multivariate DPM-G model when affine transformations are applied to the data. To this

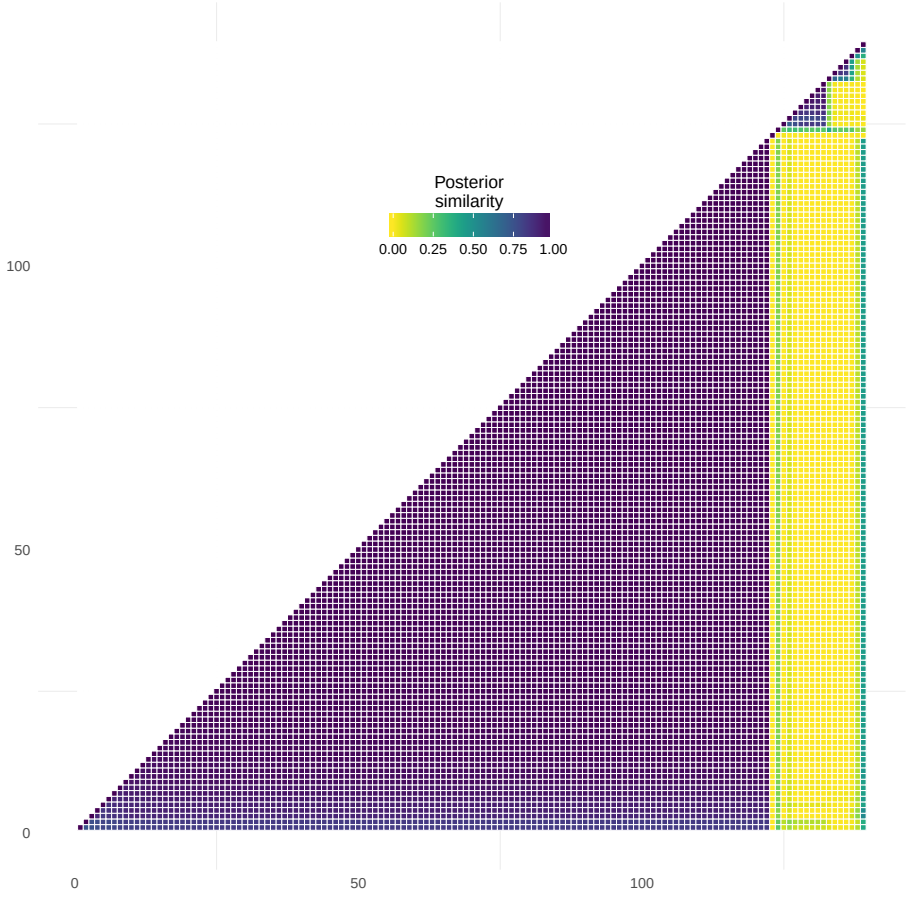

Figure 6: NGC 2419 data. Heatmap representation of the posterior similarity matrix obtained based on DPM-G model.

end we focused on the DPM-G model with independent normal and inverse-Wishart specification for the base measure. Our investigation covered both the finite sample size and the asymptotic setting. Specifically, in Proposition 1, given any affine transformation g, an explicit model specification, depending on g, was derived so to ensure coherence between posterior inferences carried out based on a dataset or its transformation via g. We then considered a different setting where the specification of the model is assumed independent of the specific transformation g. In this case, we formalised the notion of asymptotic robustness of a model under transformations of the data and showed that mild conditions on the true data generating distributions are sufficient to ensure that the DPM-G model features such a property. Specifically, Theorem 1 shows that the posterior distributions obtained conditionally on a dataset or any affine transformation of it, become more and more similar as the sample

size grows. Inference on densities and, as a by-product, on the clustering structure underlying the data, thus becomes increasingly less dependent on the affine transformation applied to the data, as the sample size grows to infinity. As a special case, Theorem 1 implies that posterior inference based DPM-G models is asymptotically robust to data transformations commonly adopted for the sake of numerical efficiency, such as standardisation or normalisation. This observation is particularly relevant when dealing with the astronomical unsupervised clustering problem motivating this work. Due to the lack of prior information on the dimensional constants relating different physical units, we resorted to a standardisation of each component of the data and chose an arbitrary model specification. Prior information was available in the form of the experts' prior opinion on the expected number of groups in the dataset and was used to elicit the hyperprior distribution for α, the total mass parameter of the DP.

Acknowledgements

The authors wish to thank Carlo Nipoti for suggesting the motivating astronomical problem and for fruitful discussions, and Éric Marchand for pointing out the example satisfying Assumption A1 but not A2 of Theorem 1. The authors are also grateful to Bob Franke for the picture in Figure 1.

Appendices

A. Proof of the lemmas

Proof of Lemma 1. We check, point-by-point, that the conditions of Theorem 2 are satisfied.

B1. Since $\boldsymbol{\mu} \sim N_d(\mathbf{m}_0, \mathbf{B}_0)$, then $\|\boldsymbol{\mu}\|^2 \sim \chi^2_d(\delta)$ where d is the dimension of $\boldsymbol{\mu}$ and $\delta = \|\mathbf{m}_0\|$ is the non-centrality parameter of the chi-squared distribution. Then, for sufficiently large x,

$$P_{0,1}\left(\|\boldsymbol{\mu}\|^2 > x\right) \leq \left(\frac{x}{d}\right)^{\frac{d}{2}} \exp\left\{\frac{d-x}{2}\right\} \lesssim x^{-2(r+1)},$$

which holds for $r > (d-1)/2$.

B2. We know that $\boldsymbol{\Sigma} \sim IW(\nu_0, \mathbf{S}_0)$ and we start by considering the case corresponding to $\mathbf{S}_0 = \mathbf{I}_d$, where $\mathbf{I}_d$ denotes the d-dimensional identity matrix. It is known that $\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}) \sim \chi^2_{\nu_0 d}$. Thus, for sufficiently large x,

$$\begin{aligned} P_{0,2}\left(\lambda_d(\boldsymbol{\Sigma}^{-1}) > x\right) &\leq P_{0,2}\left(\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}) > x\right) \\ &\leq \left(\frac{x}{\nu_0 d}\right)^{\frac{\nu_0 d}{2}} \exp\left\{\frac{\nu_0 d - x}{2}\right\} \lesssim \exp\left\{-c_1 x^{c_2}\right\}, \end{aligned}$$

for some positive constants c_1 and c_2. This result can be easily generalised to the case $\mathbf{S}_0 \neq \mathbf{I}_d$ since $IW(\mathrm{d}\boldsymbol{\Sigma}; \nu_0, \mathbf{S}_0) = \mathbf{S}_0^{-1} IW(\mathrm{d}\boldsymbol{\Sigma}; \nu_0, \mathbf{I}_d)$.

B3. We know that $\boldsymbol{\Sigma} \sim IW(\nu_0, \mathbf{S}_0)$ and we start by supposing that $\mathbf{S}_0 = \mathbf{I}_d$. The joint distribution of the eigenvalues $\boldsymbol{\lambda}\left(\boldsymbol{\Sigma}^{-1}\right)$ is known to be equal to

$$f_{\boldsymbol{\lambda}}(x_1, \ldots, x_d) = c_{d,\nu_0} \exp\left\{-\sum_{j=1}^{d} \frac{x_j}{2}\right\} \prod_{j=1}^{d} x_j^{\frac{(\nu_0 - d + 1)}{2}} \prod_{j<k} (x_k - x_j),$$

for some normalising constant c_{d,ν_0}, if $(x_1, \ldots, x_d) \in (0, \infty)^d$ is such that $x_1 \leq \cdots \leq x_d$, and equal to 0 otherwise. It is easy to verify that, on the support of $f_{\boldsymbol{\lambda}}$,

$$\prod_{j<k} (x_k - x_j) \leq \prod_{j<k} x_k = \prod_{k=2}^{d} x_k^{k-1}.$$

The density function of $\lambda_1(\boldsymbol{\Sigma}^{-1})$ then becomes

$$\begin{aligned} f_{\lambda_1}(x_1) &= \int \cdots \int f_{\boldsymbol{\lambda}}(x_1, \ldots, x_d) \mathrm{d}x_2 \cdots \mathrm{d}x_d \\ &\leq c_{d,\nu_0} x_1^{\frac{\nu_0 - d + 1}{2}} \mathrm{e}^{-\frac{x_1}{2}} \prod_{k=2}^{d} \int_0^{\infty} x_k^{\frac{\nu_0 - d + 1}{2} + k - 1} \mathrm{e}^{-\frac{x_k}{2}} \mathrm{d}x_k \end{aligned}$$

$$= c'_{d,\nu_0} x_1^{\frac{\nu_0-d+1}{2}} \exp\left\{-\frac{x_1}{2}\right\},$$

for some new normalising constant c'_{d,ν_0}. Then for any $x > 0$ we have

$$P_{0,2}\left(\lambda_1(\mathbf{\Sigma}^{-1}) < \frac{1}{x}\right) \leq c'_{d,\nu_0} \int_0^{\frac{1}{x}} x_1^{\frac{\nu_0-d+1}{2}} dx_1 \lesssim x^{-c_3 x}$$

for some constant c_3 and sufficiently large x. Again, this result can be generalised to the case $\mathbf{S}_0 \neq \mathbf{I}_d$ since $IW(\mathrm{d}\mathbf{\Sigma}; \nu_0, \mathbf{S}_0) = \mathbf{S}_0^{-1} IW(\mathrm{d}\mathbf{\Sigma}; \nu_0, \mathbf{I}_d)$.

B4. We know that $\mathbf{\Sigma} \sim IW(\nu_0, \mathbf{S}_0)$ and we start by considering the case corresponding to $\mathbf{S}_0 = \mathbf{I}_d$. We define $Z(\mathbf{\Sigma}^{-1}) = \lambda_d(\mathbf{\Sigma}^{-1})/\lambda_1(\mathbf{\Sigma}^{-1})$ and the function $q(\boldsymbol{\lambda}(\mathbf{\Sigma}^{-1})) = (\lambda_1(\mathbf{\Sigma}^{-1}), \ldots, \lambda_{d-1}(\mathbf{\Sigma}^{-1}), Z(\mathbf{\Sigma}^{-1}))$. Let $J_{q^{-1}}$ denote the Jacobian of the inverse of the function q, and observe that

$$f_{\lambda_1,\ldots,\lambda_{d-1},Z}(x_1, \ldots, x_{d-1}, z) = |J_{q^{-1}}| f_{\boldsymbol{\lambda}}(x_1, \ldots, x_{d-1}, x_1 z).$$

Then, by marginalising with respect to the first $d-1$ components, we obtain

$$\begin{aligned}
f_Z(z) &= \int \cdots \int |J_{q^{-1}}| f_{\boldsymbol{\lambda}}(x_1, \ldots, x_{d-1}, x_1 z) \mathrm{d}x_1 \cdots \mathrm{d}x_{d-1} \\
&= \int \cdots \int c_{d,\nu_0} \exp\left\{-\sum_{j=1}^{d-1} \frac{x_j}{2} - \frac{x_1 z}{2}\right\} \prod_{j=1}^{d-1} x_j^{\frac{\nu_0+1-d}{2}} (x_1 z)^{\frac{\nu_0+1-d}{2}} \\
&\quad \times \prod_{j<k\leq d-1} (x_k - x_j) \prod_{j=1}^{d-1} (x_1 z - x_j) x_1 \mathrm{d}x_1 \cdots \mathrm{d}x_{d-1} \\
&\leq \int \cdots \int c_{d,\nu_0} \exp\left\{-\sum_{j=1}^{d-1} \frac{x_j}{2} - \frac{x_1 z}{2}\right\} \prod_{j=1}^{d-1} x_j^{\frac{\nu_0+1-d}{2}} (x_1 z)^{\frac{\nu_0+1-d}{2}} \\
&\quad \prod_{k=2}^{d-1} x_k^{k-1} \prod_{j=1}^{d-1} (x_1 z) x_1 \mathrm{d}x_1 \cdots \mathrm{d}x_{d-1} \\
&\quad = c'_{d,\nu_0} z^{(\nu_0+d-1)/2} \int \exp\left\{-x_1\left(\frac{z+1}{2}\right)\right\} x_1^{\nu_0+1} \mathrm{d}x_1 \\
&\quad = c'_{d,\nu_0} (\nu_0+1)! \left(\frac{2}{z+1}\right)^{\nu_0+2} z^{(\nu_0+d-1)/2} \\
&\quad = c''_{d,\nu_0} \frac{z^{(\nu_0+d-1)/2}}{(z+1)^{\nu_0+2}} \\
&\quad \leq c''_{d,\nu_0} z^{-(\nu_0-d+5)/2},
\end{aligned}$$

for some constants c_{d,ν_0}, c'_{d,ν_0} and c''_{d,ν_0}. Thus we have

$$P_{0,2}\left(Z > x\right) = \int_x^\infty f_Z(z) \mathrm{d}z \leq c''_{d,\nu_0} \int_x^\infty z^{-(\nu_0-d+5)/2} \mathrm{d}z \lesssim x^{-\kappa},$$

for sufficiently large x, where $\kappa = (\nu_0 - d + 3)/2 > d(d+1)$ by the assumption that $\nu_0 > (d+1)(2d-3)$.

□

Proof of Lemma 2. We assume that f^* satisfies conditions A1–A4 of Theorem 1 and check that the same holds for f_g^*.

A1. Assume that $0 < f^*(\mathbf{x}) < M$ for every $\mathbf{x} \in \mathbb{R}^d$ and some $M > 0$. Then, for every $\mathbf{x} \in \mathbb{R}^d$, we have $f_g^*(\mathbf{x}) = |\det(\mathbf{C})|^{-1} f^*(g^{-1}(\mathbf{x}))$ which implies

$$0 < f_g^*(\mathbf{x}) < M' = |\det(\mathbf{C})|^{-1} M.$$

A2. Assume that f^* is such that $\left|\int f^*(\mathbf{x}) \log f^*(\mathbf{x}) \mathrm{d}\mathbf{x}\right| < \infty$. Then, we have

$$\begin{aligned}
&\left|\int f_g^*(\mathbf{x}) \log f_g^*(\mathbf{x}) \mathrm{d}\mathbf{x}\right| \\
&= \left|\int |\det(\mathbf{C})|^{-1} f^*(g^{-1}(\mathbf{x})) \log\left(|\det(\mathbf{C})|^{-1} f^*(g^{-1}(\mathbf{x}))\right) \mathrm{d}\mathbf{x}\right| \\
&= |\det(\mathbf{C})|^{-1} \left|\int f^*(g^{-1}(\mathbf{x})) \log\left(|\det(\mathbf{C})|^{-1}\right) \mathrm{d}\mathbf{x}\right. \\
&\qquad\qquad \left. + \int f^*(g^{-1}(\mathbf{x})) \log\left(f^*(g^{-1}(\mathbf{x}))\right) \mathrm{d}\mathbf{x}\right| \\
&= \left|\int f^*(\mathbf{y}) \log\left(|\det(\mathbf{C})|^{-1}\right) \mathrm{d}\mathbf{y} + \int f^*(\mathbf{y}) \log\left(f^*(\mathbf{y})\right) \mathrm{d}\mathbf{y}\right| \\
&= \left|\log\left(|\det(\mathbf{C})|^{-1}\right) + \int f^*(\mathbf{y}) \log\left(f^*(\mathbf{y})\right) \mathrm{d}\mathbf{y}\right| \\
&\leq \left|\log\left(|\det(\mathbf{C})|^{-1}\right)\right| + \left|\int f^*(\mathbf{y}) \log\left(f^*(\mathbf{y})\right) \mathrm{d}\mathbf{y}\right| < \infty.
\end{aligned}$$

A3. Assume that f^* satisfies A3 with some δ'. Let $\delta = |\det(\mathbf{C})|^{-1}\delta'$ and observe that since g is invertible

$$\begin{aligned}
\varphi_\delta^{(g)}(g(\mathbf{y})) &= \inf_{\{\mathbf{t}\,:\,\|\mathbf{t}-g(\mathbf{y})\|<\delta\}} f_g^*(\mathbf{t}) = \inf_{\{\mathbf{s}\,:\,\|g(\mathbf{s})-g(\mathbf{y})\|<\delta\}} f_g^*(g(\mathbf{s})) \\
&= \inf_{\{\mathbf{s}\,:\,\|\mathbf{s}-\mathbf{y}\|<\delta'\}} f^*(\mathbf{s}) |\det(\mathbf{C})|^{-1} = \varphi_{\delta'}(\mathbf{y}) |\det(\mathbf{C})|^{-1}.
\end{aligned}$$

Then we have that

$$\begin{aligned}
\int f_g^*(\mathbf{x}) \log\left(\frac{f_g^*(\mathbf{x})}{\varphi_\delta^{(g)}(\mathbf{x})}\right) \mathrm{d}\mathbf{x} &= \int f_g^*(g(\mathbf{y})) \log\left(\frac{f_g^*(g(\mathbf{y}))}{\varphi_\delta^{(g)}(g(\mathbf{y}))}\right) |\det(\mathbf{C})| \mathrm{d}\mathbf{y} \\
&= \int f^*(\mathbf{y}) \log\left(\frac{|\det(\mathbf{C})|^{-1} f^*(\mathbf{y})}{|\det(\mathbf{C})|^{-1} \varphi_{\delta'}(\mathbf{y})}\right) \mathrm{d}\mathbf{y}
\end{aligned}$$

$$= \int f^*(\mathbf{y}) \log\left(\frac{f^*(\mathbf{y})}{\varphi_{\delta'}(\mathbf{y})}\right) \mathrm{d}\mathbf{y} < \infty$$

where the last inequality holds by Assumption A3 on f^* with δ'. This finally shows that f_g^* satisfies Assumption A3 with δ.

A4. Observe that

$$\begin{aligned}\int \|\mathbf{x}\|^{2(1+\eta)} f_g^*(\mathbf{x})\mathrm{d}\mathbf{x} &= \int \|g(\mathbf{y})\|^{2(1+\eta)} f_g^*(g(\mathbf{y}))|\det(\mathbf{C})|\mathrm{d}\mathbf{y} \\ &= \int \|g(\mathbf{y})\|^{2(1+\eta)} f^*(\mathbf{y})\mathrm{d}\mathbf{y} \\ &\leq \int 2^{2(1+\eta)-1}\left(\|\mathbf{C}\mathbf{y}\|^{2(1+\eta)} + \|\mathbf{b}\|^{2(1+\eta)}\right) f^*(\mathbf{y})\mathrm{d}\mathbf{y},\end{aligned}$$

where the last inequality follows by combining triangular and Jensen's inequalities. Thus we can write

$$\begin{aligned}&\int \|\mathbf{x}\|^{2(1+\eta)} f_g^*(\mathbf{x})\mathrm{d}\mathbf{x} \\ &\quad \leq 2^{2(1+\eta)-1}\left(|\det(\mathbf{C})|^{2(1+\eta)} \int \|\mathbf{y}\|^{2(1+\eta)} f^*(\mathbf{y})\mathrm{d}\mathbf{y} + \|\mathbf{b}\|^{2(1+\eta)}\right) < \infty,\end{aligned}$$

where the last inequality follows by Assumption A4 on f^*.

□

References

References

Y. Ascasibar, J. Binney, Numerical estimation of densities, Monthly Notices of the Royal Astronomical Society 356 (3) (2005) 872–882.

M. Maciejewski, S. Colombi, C. Alard, F. Bouchet, C. Pichon, Phase-space structures–I. A comparison of 6D density estimators, Monthly Notices of the Royal Astronomical Society 393 (3) (2009) 703–722.

R. Ibata, A. Sollima, C. Nipoti, M. Bellazzini, S. Chapman, E. Dalessandro, The globular cluster NGC 2419: a crucible for theories of gravity, The Astrophysical Journal 738 (2) (2011) 1–23.

J. P. Huelsenbeck, P. Andolfatto, Inference of population structure under a Dirichlet process model, Genetics 175 (4) (2007) 1787–1802.

M. Medvedovic, S. Sivaganesan, Bayesian infinite mixture model based clustering of gene expression profiles, Bioinformatics 18 (9) (2002) 1194–1206.

Y. Xu, P. F. Thall, P. Müller, R. J. Mehran, A decision-theoretic comparison of treatments to resolve air leaks after lung surgery based on nonparametric modeling, Bayesian Analysis 12 (3) (2017) 639–652.

E. Otranto, G. M. Gallo, A nonparametric Bayesian approach to detect the number of regimes in Markov switching models, Econometric Reviews 21 (4) (2002) 477–496.

A. Y. Lo, On a class of Bayesian nonparametric estimates: I. Density estimates, The Annals of Statistics 12 (1) (1984) 351–357.

N. L. Hjort, C. Holmes, P. Müller, S. G. Walker, Bayesian nonparametrics, vol. 28, Cambridge University Press, 2010.

J. Arbel, B. Nipoti, Discussion of "Bayesian Nonparametric Inference Why and How Comment", by Müller and Mitra, Bayesian Analysis 8 (02) (2013) 326–328.

A. Bean, X. Xu, S. MacEachern, Transformations and Bayesian density estimation, Electronic Journal of Statistics 10 (2) (2016) 3355–3373.

P. Müller, A. Erkanli, M. West, Bayesian curve fitting using multivariate normal mixtures, Biometrika 83 (1) (1996) 67–79.

Y. Wu, S. Ghosal, The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation, Journal of Multivariate Analysis 101 (10) (2010) 2411 – 2419, ISSN 0047-259X.

W. Shen, S. T. Tokdar, S. Ghosal, Adaptive Bayesian multivariate density estimation with Dirichlet mixtures, Biometrika 100 (3) (2013) 623–640.

A. Canale, P. De Blasi, Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation, Bernoulli 23 (1) (2017) 379–404.

T. Ferguson, A Bayesian analysis of some nonparametric problems, The Annals of Statistics 1 (2) (1973) 209–230, ISSN 0090-5364.

J. Sethuraman, A constructive definition of Dirichlet priors, Statistica Sinica 4 (1994) 639–650.

A. Canale, B. Scarpa, Bayesian nonparametric location–scale–shape mixtures, Test 25 (1) (2016) 113–130.

A. Kottas, Nonparametric Bayesian survival analysis using mixtures of Weibull distributions, Journal of Statistical Planning and Inference 136 (3) (2006) 578–596.

M. Krnjajić, A. Kottas, D. Draper, Parametric and nonparametric Bayesian model specification: A case study involving models for count data, Computational Statistics & Data Analysis 52 (4) (2008) 2110–2128.

M. D. Escobar, M. West, Bayesian density estimation and inference using mixtures, Journal of the American Statistical Association 90 (430) (1995) 577–588.

D. Görür, C. E. Rasmussen, Dirichlet process gaussian mixture models: Choice of the base distribution, Journal of Computer Science and Technology 25 (4) (2010) 653–664.

E. L. Lehmann, G. Casella, Theory of point estimation, Springer Science & Business Media, 2006.

S. Petrone, J. Rousseau, C. Scricciolo, Bayes and empirical Bayes: do they merge?, Biometrika 101 (2) (2014) 285–302.

S. Donnet, V. Rivoirard, J. Rousseau, C. Scricciolo, Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures, Bernoulli 24 (1) (2018) 231–256.

J. Rousseau, On the frequentist properties of Bayesian nonparametric methods, Annual Review of Statistics and Its Application 3 (2016) 211–231.

S. Wade, Z. Ghahramani, Bayesian Cluster Analysis: Point Estimation and Credible Balls, Bayesian Analysis (2018) 1–29.