ISSN 2279-9362

Research
Education
Outreach

CCA

Nonparametric priors with full-range borrowing of information

Filippo Ascolani, Beatrice Franzolini, Antonio Lijoi and Igor Prunster

No. 705

December 2023

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Nonparametric priors with full-range borrowing of information

Filippo Ascolani[1,2], Beatrice Franzolini[1], Antonio Lijoi[1,2], and Igor Prünster[1,2]

[1] Bocconi Institute for Data Science and Analytics (BIDSA), Bocconi University Milano, Italy

[2]Collegio Carlo Alberto, Torino, Italy

Abstract

Modeling of the dependence structure across heterogeneous data is crucial for Bayesian inference since it directly impacts the borrowing of information. Despite the extensive advances over the last two decades, most available proposals allow only for non–negative correlations. We derive a new class of dependent nonparametric priors that can induce correlations of any sign, thus introducing a new and more flexible idea of borrowing of information. This is achieved thanks to a novel concept, which we term *hyper-tie*, and represents a direct and simple measure of dependence. We investigate prior and posterior distributional properties of the model and develop algorithms to perform posterior inference. Illustrative examples on simulated and real data show that our proposal outperforms alternatives in terms of prediction and clustering.

Keywords: Bayesian nonparametrics; Borrowing of information; Completely random measure; Dependent nonparametric prior; Negative correlation; Partial exchangeability.

1 Introduction

Bayesian nonparametric methods are increasingly popular, mainly thanks to their flexibility and strong foundations. The most common assumption underlying Bayesian models is exchangeability, which corresponds to invariance of the joint distribution of the observations with respect to finite permutations. However, real phenomena often present a level of heterogeneity that makes exchangeability unrealistic: collected data may refer to different features, populations, or, in general, may be collected under different experimental conditions. Such situations entail a significant level of heterogeneity and opportunities for borrowing information, that can be exploited through the notion of

partial exchangeability, which implies exchangeability within each experimental condition, but not across. Two sequences of observations $X = (X_i)_{i\geq 1}$ and $Y = (Y_j)_{j\geq 1}$, taking values in a space $\mathbb{X}$, are partially exchangeable if and only if, for all sample sizes (n, m) and all permutations (π_1, π_2),

$$\big((X_i)_{i=1}^n, (Y_j)_{j=1}^m\big) \stackrel{d}{=} \big((X_{\pi_1(i)})_{i=1}^n, (Y_{\pi_2(j)})_{j=1}^m\big).$$

with $\stackrel{d}{=}$ denoting equality in distribution. From an inferential point of view, partial exchangeability entails that the order of the observations within each sample is non-informative, while the belonging to a specific sample is relevant and has to be taken into account. Moreover, by de Finetti's representation theorem (de Finetti, 1938) X and Y are partially exchangeable if and only if there exist random probabilities $(\tilde{p}_1, \tilde{p}_2)$ such that for any $i, j = 1, \ldots, n$

$$(X_i, Y_j) \mid (\tilde{p}_1, \tilde{p}_2) \overset{iid}{\sim} \tilde{p}_1 \times \tilde{p}_2 \qquad (\tilde{p}_1,\, \tilde{p}_2) \sim Q \tag{1}$$

with Q playing the role of the prior. The dependence induced by Q at the level of the observables defines the Bayesian learning mechanism and it connects to the notion of borrowing of information. This term was first coined by John Tukey (Brillinger, 2002) and popularized with reference to Stein's paradox and empirical Bayes techniques in Efron and Morris (1977). More generally, statisticians refer to borrowing of information when many samples contribute to inference related to just one sample. Imagine collecting the samples $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$, while being interested only in the parameter $\tilde{p}_1$ associated to X. The simplest approach could be to disregard the second sample $(Y_j)_{j=1}^m$, with the drawback of losing potentially useful information. The typical borrowing instead consists in shrinking the estimates for different samples towards each other: shrinkage is justified by the fact that distributions of different, but related, populations are expected to be similar in terms of shape and/or location. However, many contexts may still require borrowing of information between $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$, but without necessarily resulting in shrinkage. Indeed, one's available prior information may imply that the responses in different groups have a negative association and, thus, tend to be dissimilar in location, which makes shrinkage undesirable. Similarly, when there is no pre-experimental knowledge on the dependence between X_i and Y_j, a flexible prior specification allowing also for negative association would be more appropriate. A toy parametric example to further clarify that borrowing does not necessarily imply classic shrinkage is provided in Section S2. of the Appendix. Some applied scenarios of borrowing of information not resulting in shrinkage are, for instance, the study of survival times and abundances of competitive species (Lee et al., 2020), the incorporation of retrospective data to study associations between biomarkers (Gong et al., 2021), the association between dental caries and dental fluorosis (Lorenz et al., 2018), the analysis of stocks and bonds returns (see Bhardwaj and Dunsby, 2013, and Section 6.3), and the clustering of multivariate responses with missing entries (see Section 6.4). In this paper we introduce a class of nonparametric

priors that allows for a more general version of borrowing, which includes shrinkage as a special case. These can be used as core building blocks for models tailored to specific applications.

Starting from the pioneering works of Cifarelli and Regazzini (1978) and MacEachern (1999, 2000), Bayesian nonparametric contributions for non–exchangeable data have grown substantially, see Foti and Williamson (2013), Müller et al. (2015) and Quintana et al. (2022) for insightful reviews. The vast majority of nonparametric models for partially exchangeable data entails that the random probabilities in (12) are such that

$$\begin{cases} \tilde{p}_1 \overset{a.s.}{=} \sum_{k\geq 1} \bar{J}_k \delta_{\theta_k} \\ \tilde{p}_2 \overset{a.s.}{=} \sum_{k\geq 1} \bar{W}_k \delta_{\phi_k} \end{cases} \qquad \theta_k \overset{\text{i.i.d.}}{\sim} P_0, \quad \phi_k \overset{\text{i.i.d.}}{\sim} P_0 \tag{2}$$

where the random weights $\big((\bar{J}_k),(\bar{W}_k)\big)$ and the atoms $((\theta_k),(\phi_k))$ are independent and $\theta_k \perp \phi_h$ for $k \neq h$. In this paper we focus on this class of models and, for ease of exposition, take $\tilde{p}_1$ and $\tilde{p}_2$ with the same marginal distribution.

A first prominent strategy for defining Q is to explicitly assign the distribution of the weights and the atoms in (2) so to create dependence between $\tilde{p}_1$ and $\tilde{p}_2$: this approach has led to dependent Dirichlet processes (MacEachern, 1999, 2000; Quintana et al., 2022), dependent stick-breaking processes, kernel stick-breaking processes (Dunson and Park, 2008), probit stick-breaking processes (Rodriguez and Dunson, 2011) and others. Despite their flexibility and the availability of posterior sampling schemes, the derivation of analytical results is very difficult for these models; it is often not clear how the dependence of the series reflects at the level of the observables and therefore such methods may lack transparency.

A second popular strategy, analytically more tractable, relies on completely random measures (CRMs) either working directly on the law of multi-dimensional vectors of CRMs (Epifani and Lijoi, 2010; Griffin and Leisen, 2017; Riva-Palacio and Leisen, 2021) or combining conditionally independent CRMs, using additive structures (Müller et al., 2004; Griffin et al., 2013; Lijoi and Nipoti, 2014; Lijoi et al., 2014a,b), nested structures (Rodriguez et al., 2008; Camerlenghi et al., 2019), or hierarchical structures (Teh et al., 2006; Camerlenghi et al., 2019). CRMs are then suitably transformed to obtain the random probabilities in (2).

Dependent random probabilities clearly induce dependence across groups of observations. The simplest and most intuitive way to quantify the dependence structure is through correlations. Therefore, when considering correlations among observables, we will implicitly assume real-valued X_i's and Y_j's, namely $\mathbb{X} = \mathbb{R}$. All other results and concepts are valid for general spaces $\mathbb{X}$. A first result in this direction shows that, regardless of the specific dependent model, observations in different

groups cannot be more correlated (in absolute sense) than the ones in the same group.

Proposition 1. *Suppose X and Y are partially exchangeable sequences, such that $\tilde{p}_1$ and $\tilde{p}_2$ in* (12) *have the same marginal distribution. Then*

$$-\text{corr}(X_i, X_{i'}) \leq \text{corr}(X_i, Y_j) \leq \text{corr}(X_i, X_{i'}),$$

for any i, i' and j.

Due to exchangeability within each group, the upper bound in Proposition 1 is always non–negative and it can be shown that, for all the models as in (2), the correlation between observations in the same sample, $\text{corr}(X_i, X_{i'})$, is determined by the probability of a tie. As for the correlation across samples $\text{corr}(X_i, Y_j)$, we show that a similar result holds true, with *hyper-ties*, the new notion we introduce, replacing ties.

Moreover, note that for most models based on CRMs, which allow for the computation of the correlation, $\text{corr}(X_i, Y_j)$ turns out to be positive: this happens in particular when the interaction between two or more groups is of interest. Therefore, the literature available to date within the partially exchangeable setting is focused on models that attain a limited range of possible values of the correlation, when it can be evaluated. Here we aim to overcome this limitation and introduce a novel class of priors which yield a wider range of correlation values among the observables, including those with negative sign. The next result shows that the sign of the correlation is only determined by the dependence structure between the atoms.

Proposition 2. *Suppose X and Y are partially exchangeable sequences, such that the underlying $\tilde{p}_1$ and $\tilde{p}_2$ are as in* (2)*. Moreover, for any k and k', let* $\text{corr}(\theta_k, \phi_{k'}) \geq 0$*. Then* $\text{corr}\,(X_i, Y_j) \geq 0$*, for any i and j.*

For instance, hierarchical processes (Teh et al., 2006; Camerlenghi et al., 2019), which represent one of the most popular dependent models, induce dependence by the sharing of atoms across groups. However, by Proposition 2, this means that achieving negative correlation is impossible. Hence, a flexible joint distribution for the sequence of atoms must be specified. This task is accomplished by our proposal, termed normalized CRMs with Full-Range Borrowing of Information (n-FuRBI), that allows to attain any possible value for the correlation specified in Proposition 1. Moreover, it encompasses many previous constructions as special cases. We will show that it nicely combines the flexibility of the random series construction with the analytical tractability featured by CRMs. Our proposal allows to consider any interesting choice of borrowing of information: independence, classical shrinkage, but also repulsion of estimates for different samples, generating what we term *full–range borrowing of information*. Note that the repulsive behaviour of n-FuRBI is different from

the one featured by the priors introduced in Petralia et al. (2012) and Quinlan et al. (2017), that induce repulsion among the atoms of a single random probability measure.

The appendix includes all the analytical derivations and proofs, the simulation algorithms for the implementation of the proposed class of models, additional examples and numerical studies. In the following we use the prefix S to indicate sections of the Appendix. The code to allow full replication of the numerical results is available at `https://github.com/beatricefranzolini/FuRBI`.

2 General results on dependent processes

The vast majority of dependent processes introduced in the literature are almost surely discrete and therefore admit a series representation as in (2). A key preliminary step leading to the definition of hyper-tie and n-FuRBI priors is the observation that the random probabilities in (2) can be embedded into

$$\begin{cases} p_1 \overset{a.s.}{=} \sum_{k\geq 1} \bar{J}_k \delta_{(\theta_k,\phi_k)} \\ p_2 \overset{a.s.}{=} \sum_{k\geq 1} \bar{W}_k \delta_{(\theta_k,\phi_k)} \end{cases} \qquad (\theta_k, \phi_k) \overset{i.i.d.}{\sim} G_0, \tag{3}$$

with G_0 a probability distribution on $\mathbb{X} \times \mathbb{X}$, whose marginals equal P_0. While p_1 and p_2 share the same atoms, the weights and the atoms are independent and the pair of random probability measures $\tilde{p}_1$ and $\tilde{p}_2$ in (2) are obtained as the projections over different coordinates of p_1 and p_2, namely $\tilde{p}_1(\cdot) = p_1(\cdot \times \mathbb{X})$ and $\tilde{p}_2(\cdot) = p_2(\mathbb{X} \times \cdot)$. The structure of popular models is recovered by letting either $G_0 = P_0^2$, which corresponds to independence, or $G_0(\mathrm{d}\theta, \mathrm{d}\phi) = P_0(\mathrm{d}\theta)\delta_{\{\theta\}}(\mathrm{d}\phi)$, that is $\theta_k = \phi_k$ for any k as happens for, e.g., hierarchical processes (see Camerlenghi et al., 2019). Almost sure discreteness implies that a sample from the random probability measure $\tilde{p}_1$ (or $\tilde{p}_2$) will display ties with positive probability. The probability of a tie, i.e. a coincidence of any two observations i and j in the same sample, is

$$\beta := \mathbb{P}(X_i = X_j) = \sum_{k\geq 1} \mathrm{E}(\bar{J}_k^2) = \sum_{k\geq 1} \mathrm{E}(\bar{W}_k^2) = \mathbb{P}(Y_i = Y_j) \tag{4}$$

with $(\bar{J}_k)_{k\geq 1}$ and $(\bar{W}_k)_{k\geq 1}$ equal in distribution since we are assuming, for simplicity, that $\tilde{p}_1$ and $\tilde{p}_2$ are equal in distribution. When considering jointly the two samples, the concept of tie can be replaced by the one of *hyper-tie*, that is two observations in different samples coinciding with components having the same label. According to (12), its probability is

$$\gamma := \sum_{k\geq 1} \mathbb{P}(X_i = \theta_k,\, Y_j = \phi_k) = \sum_{k\geq 1} \mathrm{E}(\bar{J}_k \bar{W}_k). \tag{5}$$

Sampling from components with the same label is equivalent to sampling the same atom at the level of the underlying (p_1, p_2) in (3). Clearly, when the atoms are shared between $\tilde{p}_1$ and $\tilde{p}_2$, i.e. $G_0(\mathrm{d}\theta, \mathrm{d}\phi) = P_0(\mathrm{d}\theta)\delta_{\{\theta\}}(\mathrm{d}\phi)$, a hyper-tie corresponds to an actual tie between observations in different samples.

The next result shows the relationship between β and γ, the probabilities of a tie and hyper-tie, respectively: in particular, the probability of a tie is always larger and equality is attained if and only if the probability masses of p_1 and p_2 are almost surely equal.

Proposition 3. *Let* $(\tilde{p}_1, \tilde{p}_2)$ *be as in* (2) *and* β, γ *as in* (4) *and* (5)*, respectively. Then* $0 \leq \gamma \leq \beta$ *and* $\beta = \gamma$ *if and only if* $\bar{W}_k \overset{a.s.}{=} \bar{J}_k$ *for any* k*.*

Hyper-ties play a crucial role in determining the dependence between observables across groups, as the ties do for the dependence between observables within groups, as shown by the next proposition.

Proposition 4. *Consider model* (12) *with* $(\tilde{p}_1, \tilde{p}_2)$ *as in* (2)*. Then, for any* $i \neq i'$ *and any* $j \neq j'$

$$\text{corr}(X_i, X_{i'}) = \text{corr}(Y_j, Y_{j'}) = \beta \qquad\qquad \text{corr}(X_i, Y_j) = \gamma\, \rho_0$$

with ρ_0 *the correlation between two random variables jointly sampled from* G_0*.*

Thus, while the correlation between observations in the same sample equals the probability of a tie, the correlation between observations from different samples is determined by the probability of a hyper-tie, corrected by the correlation between atoms. Clearly a suitable choice of the joint distribution of the atoms makes the latter negative. Thus, by choosing G_0 appropriately, for instance as a bivariate normal, it is easy to tune the correlation according to the available prior knowledge. The following Corollary shows the values that can be attained, once the marginal law is specified.

Corollary 1. *Consider model* (12) *with* $(\tilde{p}_1, \tilde{p}_2)$ *as in* (2)*. If the marginal distribution of* $\tilde{p}_1$ *and* $\tilde{p}_2$ *is fixed, then* $\text{corr}(X_i, Y_j) \in [-\beta, \beta]$ *and the extreme values are attained if and only if the jumps are equal and* $\rho_0 = \pm 1$*.*

Unsurprisingly, with equal weights and jumps, which corresponds to full exchangeability, one achieves the extreme case of $\text{corr}(X_i, Y_j) = \beta$. Null correlation, instead, is attained when atoms are uncorrelated or when the probability of hyper-ties is zero. Lastly, maximum negative correlation $\text{corr}(X_i, Y_j) = -\beta$ is attained with equal weights and negatively correlated atoms and can be thought of as the opposite case with respect to exchangeability, at least in terms of correlation. Ties and hyper-ties play a similar role also in the predictive structure, as the next result shows.

Proposition 5. *Consider model* (12) *with* $(\tilde{p}_1, \tilde{p}_2)$ *as in* (2)*. Then*

$$\mathbb{P}\left(X_1 \in A, X_2 \in B\right) = \beta P_0(A \cap B) + (1-\beta)\, P_0(A) P_0(B).$$

and

$$\mathbb{P}\left(X_1 \in A, Y_1 \in B\right) = \gamma G_0(A \times B) + (1-\gamma)\, P_0(A) P_0(B).$$

The result is indeed quite intuitive. If X_1 and Y_1 form a hyper-tie (with probability γ) they come from the same pair of atoms and need to be sampled jointly; otherwise they refer to different atoms and are sampled independently. The same happens inside each group, where X_1 and X_2 are equal with probability β.

Example 1. The hierarchical Dirichlet process (Teh et al., 2006) is characterized by the hierarchical representation $\tilde{p}_i \mid \tilde{p}_0 \overset{\text{i.i.d.}}{\sim} \text{DP}(\theta, \tilde{p}_0)$, with $\tilde{p}_0 \sim \text{DP}(\theta_0, P_0)$, where P_0 is a diffuse measure and $\text{DP}(\alpha, H)$ denotes the law of a Dirichlet process with concentration parameter $\alpha > 0$ and baseline distribution H. Since the $\tilde{p}_i$'s share the atoms, an hyper-tie corresponds to an actual tie between observations in different samples, so that with simple computations we get

$$\beta = \text{corr}(X_i, X_j) = 1 - \frac{\theta\theta_0}{(1+\theta)(1+\theta_0)}, \qquad \gamma = \text{corr}(X_i, Y_j) = \frac{1}{1+\theta_0}.$$

Thus, the correlation among the observables is forced to be positive, with θ_0 tuning the dependence; see Example 1 in Camerlenghi et al. (2019) for more details.

Given the above results and considerations, it should be clear that γ defined in (5) is crucial for tuning the level of dependence. However, closed form expressions of γ are available only for a few cases and, in fact, we are facing a trade–off: on the one hand we have dependent processes based on the stick-breaking representation, that allow for high flexibility while sacrificing the availability of analytical results; on the other hand we have constructions based on CRMs, for which an extensive theory has been developed, though they are not as effective for tuning the dependence, since all the existing instances produce non-negative correlation across samples. In the following we combine the best of both approaches through n-FuRBI: they are flexible processes that can attain any value for the correlation between the observables, while at the same time a posterior representation can be derived. Their construction is based on CRMs and completely random vectors, reviewed in the next section.

3 Some basics on completely random measures

As shown in Lijoi and Prünster (2010), many Bayesian nonparametric models can be obtained as suitable transformations of CRMs; among others, these include the Dirichlet process, the Pitman-Yor process and the neutral-to-the-right priors. The extension of CRMs to the bivariate setting is provided by *completely random vectors* $\mu = (\mu_1, \mu_2)$, whose components take values in the space of boundedly finite measures on $\mathbb{X}$ and are such that, for every collection of pairwise disjoint sets $(A_i)_{i\geq 1}^n$, the random vectors $(\mu_1(A_1), \mu_2(A_1)), \ldots, (\mu_1(A_n), \mu_2(A_n))$ are mutually independent. We focus on the case of no fixed atoms and no deterministic component, so that the marginal CRMs μ_1 and μ_2 are almost surely discrete and can be written as sum of $\mathbb{X}$–valued random atoms with random weights, i.e.

$$\mu_1 \overset{a.s.}{=} \sum_{i\geq 1} J_i \delta_{\kappa_i}, \quad \mu_2 \overset{a.s.}{=} \sum_{i\geq 1} W_i \delta_{\kappa_i}.$$

In the following section it will be convenient to use the reparametrization $\kappa_i = (\theta_i, \phi_i) \in \mathbb{X} = \mathbb{X}_1 \times \mathbb{X}_2$. Such completely random vectors are characterized by the Lévy-Khintchine representation

$$\mathbb{E}\left\{e^{-\mu_1(f_1)-\mu_2(f_2)}\right\} = \exp\left[-\int_{\mathbb{R}_+^2\times\mathbb{X}} \{1 - e^{-s_1 f_1(x) - s_2 f_2(x)}\}\, v(\mathrm{d}s_1, \mathrm{d}s_2, \mathrm{d}x)\right] \tag{6}$$

where $\mu_i(f_i) = \int_{\mathbb{X}} f_i(x)\mu_i(\mathrm{d}x)$ for $\mathbb{R}^+$-valued f_i and $v(\mathrm{d}s_1, \mathrm{d}s_2, \mathrm{d}x)$ is the joint Lévy intensity. We shall focus on the homogeneous case, in which jumps $(J_j)_{j\geq 1}$ and locations $(X_j)_{j\geq 1}$ are independent. In terms of Lévy intensity it reads $v(\mathrm{d}s_1, \mathrm{d}s_2, \mathrm{d}x) = \rho(\mathrm{d}s_1, \mathrm{d}s_2)\alpha(\mathrm{d}x)$ for some finite measure α on $\mathbb{X}$ and measure ρ. Moreover, in the sequel we will also need the joint and marginal Laplace exponents given by

$$\psi_b(\lambda_1, \lambda_2) := \int_{\mathbb{R}_+^2\times\mathbb{X}} (1 - e^{-\lambda_1 s_1 - \lambda_2 s_2})\rho(\mathrm{d}s_1, \mathrm{d}s_2)\alpha(\mathrm{d}x), \quad \lambda_1 > 0, \lambda_2 > 0.$$

$$\psi(\lambda) := \int_{\mathbb{R}_+\times\mathbb{X}} (1 - e^{-\lambda s})\rho(\mathrm{d}s)\alpha(\mathrm{d}x) \quad \lambda > 0,$$

For an exhaustive account on CRMs, we refer to Kingman (1967, 1993). Completely random vectors and CRMs are often normalized to obtain random probability measures, as introduced in Regazzini et al. (2003), i.e. $p(\cdot) = \mu(\cdot)/\mu(\mathbb{X})$. Notice that in principle any random measure μ such that $\mathbb{P}(0 < \mu(\mathbb{X}) < \infty) = 1$ can be normalized in order to define a random probability measure. However, the strength of completely random vectors and measures lies in their Lévy–Khintchine representations and unique correspondence with the associated Lévy intensity, which allow a high

degree of analytical tractability. CRMs and the corresponding normalized probabilities have been extensively studied to model exchangeable data (see, for instance, James et al., 2006, 2009, 2010; Lijoi and Prünster, 2010; Favaro et al., 2016; Camerlenghi et al., 2018). Similarly, a completely random vector can be used to model dependence between two groups. For more details on completely random vectors and an interesting account of their dependence structure, we refer to Catalano et al. (2021, 2023). Since the two measures in the vector share all the atoms, by virtue of Proposition 2 the induced model yields non–negative correlation between samples. The issue is addressed in the next section, by means of a novel class of random probability measures that leverage the dependence structure specifed for the atoms.

4 Full-range borrowing of information nonparametric prior

4.1 Definition and first properties

In this section we introduce n-FuRBI and for simplicity we still consider only the case of two samples with the same a priori marginal distribution.

Definition 1. Consider a completely random vector (μ_1, μ_2) on $\mathbb{X}^2$ with Lévy intensity

$$v(\mathrm{d}s_1, \mathrm{d}s_2, \mathrm{d}x_1, \mathrm{d}x_2) = \rho(\mathrm{d}s_1, \mathrm{d}s_2)\, \alpha(\mathrm{d}x_1, \mathrm{d}x_2),$$

where $\alpha(\mathrm{d}x_1, \mathrm{d}x_2) = \theta G_0(\mathrm{d}x_1, \mathrm{d}x_2)$, where $\theta = \alpha(\mathbb{X}^2) \in (0, +\infty)$, and G_0 is a non-atomic probability measure on $\mathbb{X}^2$ such that $G_0(\cdot \times \mathbb{X}) = G_0(\mathbb{X} \times \cdot) = P_0(\cdot)$. Then $\tilde{\mu}_1$ and $\tilde{\mu}_2$ defined as

$$\tilde{\mu}_1(\cdot) = \mu_1(\mathbb{X} \times \cdot) \qquad \tilde{\mu}_2(\cdot) = \mu_2(\cdot \times \mathbb{X})$$

are CRMs with *Full-Range Borrowing of Information* (FuRBI CRMs) and underlying Lévy intensity v. The normalized versions $\tilde{p}_j(\cdot) = \tilde{\mu}_j(\cdot)/\tilde{\mu}_j(\mathbb{X})$ for $j = 1, 2$ are said *normalized CRMs with Full-Range Borrowing of Information* (n-FuRBI).

Essentially, first a pair of random measures endowed with the same locations is constructed on the product space $\mathbb{X}^2$; as a second step, the coordinates of each pair of atoms are split. Thus, the n-FuRBI admit a representation as in (2) and (3). In general FuRBI CRMs are not completely random vectors, because the joint sampling of the atoms forbids the independence of the vector evaluated on pairwise disjoint sets. However, the representation in terms of a completely random vector in the product space is useful to characterize the joint law of the FuRBI CRMs, as shown in the following proposition.

Proposition 6. *Let $(\tilde{\mu}_1, \tilde{\mu}_2)$ be a vector of FuRBI CRMs. Then*

(i) *$\tilde{\mu}_1$ and $\tilde{\mu}_2$ are CRMs with intensity $\rho(\mathrm{d}s)\theta P_0(\mathrm{d}x)$, where $\rho(\mathrm{d}s) = \int_{\mathbb{R}_+} \rho(\mathrm{d}s_1, \mathrm{d}s)$.*

(ii) *For any A and B, the following equality holds*

$$E\big[\mathrm{e}^{-\lambda_1\tilde{\mu}_1(A)-\lambda_2\tilde{\mu}_2(B)}\big] = \exp\{-G_0(A\times B^c)\psi(\lambda_1) - G_0(A^c\times B)\psi(\lambda_2)\} \times \exp\{-G_0(A\times B)\psi_b(\lambda_1,\lambda_2)\},$$

where ψ denotes the common marginal Laplace exponent and ψ_b the joint Laplace exponent of (μ_1, μ_2).

(iii) *The joint law of $(\tilde{\mu}_1, \tilde{\mu}_2)$ is characterized by the joint Lévy intensity of (μ_1, μ_2).*

The next proposition shows that the β and γ associated to any couple of n-FuRBI can be computed through their Laplace exponents.

Proposition 7. *Consider $(\tilde{p}_1, \tilde{p}_2)$ n-FuRBI. Then the probability of a tie and of a hyper-tie are respectively*

$$\beta = -\int_{\mathbb{R}_+} u\left\{\frac{\mathrm{d}^2}{\mathrm{d}u^2}\psi(u)\right\} e^{-\psi(u)}\,\mathrm{d}u, \quad \gamma = -\int_{\mathbb{R}_+^2}\left\{\frac{\partial^2}{\partial u_1\partial u_2}\psi_b(u_1,u_2)\right\} e^{-\psi_b(u_1,u_2)}\,\mathrm{d}u_1\mathrm{d}u_2.$$

Thus, the crucial value of γ can be obtained by computing, analytically or numerically, a bivariate integral. The two results above show a recurrent trait of our approach: interesting quantities will be usually rewritten in terms of the original completely random vector, in order to exploit its analytical tractability. We conclude this section with two examples of FuRBI CRMs, that also show how some existing constructions can be obtained as special cases.

Example 2 (FuRBI CRMs with equal jumps)**.** Let $\rho(\mathrm{d}s_1)\delta_{s_1}(\mathrm{d}s_2)\,\theta\,G_0(\mathrm{d}x_1,\mathrm{d}x_2)$ be the underlying Lévy intensity. The series representation of the corresponding FuRBI CRMs is

$$\tilde{\mu}_1(\cdot) \overset{a.s.}{=} \sum_{k\geq 1} W_k\delta_{\theta_k} \qquad \tilde{\mu}_2(\cdot) \overset{a.s.}{=} \sum_{k\geq 1} W_k\delta_{\phi_k} \qquad \text{with } (\theta_k,\phi_k) \overset{i.i.d}{\sim} G_0.$$

Therefore, $\gamma = \beta$, so that a tie and a hyper–tie are observed with the same probability.

Example 3 (Extended Compound FuRBI CRMs)**.** Consider the Lévy intensity

$$v(\mathrm{d}s_1,\mathrm{d}s_2,\mathrm{d}x_1,\mathrm{d}x_2) = \int z^{-2}h(s_1/z,\ s_2/z)\,\mathrm{d}s_1\mathrm{d}s_2v^*(\mathrm{d}z)\,\theta\,G_0(\mathrm{d}x_1,\mathrm{d}x_2),$$

where h is some density and v^* is a Lévy intensity that satisfies

$$\int z^{-2} \int \min\{1, ||s||\} h(s_1/z,\, s_2/z)\, \mathrm{d}s_1 \mathrm{d}s_2 v^*(\mathrm{d}z) < \infty, \quad ||s|| = \sqrt{s_1^2 + s_2^2}.$$

The series representation of the corresponding FuRBI CRMs is

$$\tilde{\mu}_1(\cdot) \overset{a.s.}{=} \sum_{k\geq 1} m_{1,k} W_k \delta_{\theta_k} \qquad\qquad \tilde{\mu}_2(\cdot) \overset{a.s.}{=} \sum_{k\geq 1} m_{2,k} W_k \delta_{\phi_k}$$

where $(\theta_k, \phi_k) \overset{i.i.d}{\sim} G_0$ and $(m_{1,k}, m_{2,k}) \overset{iid}{\sim} h$. When G_0 is degenerate on the main diagonal, one retrieves the class of compound random measures introduced by Griffin and Leisen (2017).

4.2 Correlation structure between n-FuRBI

In order to analyze the dependence between the marginal n-FuRBI priors $\tilde{p}_1$ and $\tilde{p}_2$, it is useful to compute the correlation of the random probability measures evaluated on the same set A. In all the existing CRM-based models such a correlation does not depend on the specific set considered and, hence, it is often used as a global measure of dependence. The next proposition provides the covariance structure between two n-FuRBI.

Proposition 8. *Let $\tilde{p}_1$ and $\tilde{p}_2$ be n-FuRBI. Then for any A, B, such that $0 \leq P_0(A) \leq 1$ and $0 \leq P_0(B) \leq 1$, we have* $\mathrm{cov}(\tilde{p}_1(A), \tilde{p}_2(B)) = \gamma\,[G_0(A \times B) - P_0(A)P_0(B)]$ *and*

$$\mathrm{corr}(\tilde{p}_1(A), \tilde{p}_2(B)) = \frac{\gamma}{\beta} \frac{G_0(A \times B) - P_0(A)P_0(B)}{\sqrt{P_0(A)(1 - P_0(A))P_0(B)(1 - P_0(B))}}.$$

By setting $A = B$, from the previous results one immediately deduces that $\mathrm{cov}(\tilde{p}_1(A), \tilde{p}_2(A)) = \gamma\,[G_0(A \times A) - P_0(A)^2]$ and

$$\mathrm{corr}(\tilde{p}_1(A), \tilde{p}_2(A)) = \frac{\gamma}{\beta} \frac{G_0(A \times A) - P_0(A)^2}{P_0(A)(1 - P_0(A))}.$$

Unlike what usually happens with existing models, here the correlation can be negative, when A is such that $G_0(A \times A) < P_0(A)^2$, that is when G_0 exhibits a repulsive behaviour between the coordinates in $\mathbb{X}^2$. Moreover, the correlation depends on the specific set on which the two measures are evaluated and, therefore, it has to be interpreted as a local measure of dependence. See Section S3. for an illustration of this phenomenon on sets of the form $(-\infty, x)$.

Example 4 (n-FuRBI with equal jumps)**.** In this case, Proposition 3 entails $\beta = \gamma$. Therefore

$$\text{corr}\left(\tilde{p}_1(A), \tilde{p}_2(A)\right) = \frac{G_0(A \times A) - P_0(A)^2}{P_0(A)(1 - P_0(A))}.$$

Moreover, still by virtue of Proposition 3, for a given G_0 this is the highest possible correlation in absolute value.

Proposition 4 then provides the correlation between the observables, which is even more important from a modeling perspective.

Example 5 (Gamma n-FuRBI with equal jumps)**.** If the common marginal is the law of a Dirichlet process, then $\text{corr}(X_i, Y_j) = \rho_0/(1 + \theta)$. Choosing appropriately ρ_0 and θ the entire range $(-1, 1)$ becomes available.

Note that hyper-ties allow to perform a more general type of borrowing, compared to ties, even when the correlation is positive. While ties are a useful construction to model multiple samples that share certain values/latent parameters, hyper-ties can borrow information even when the two samples have no common values/latent parameter. This aspect will play a crucial role in the data-analyses of Sections 6.3 and 6.4; for these the assumption of common values would be highly unrealistic.

5 Inference

5.1 Posterior Characterization

Having provided an exhaustive description of the a priori properties of n-FuRBI, the following key step is to provide a tractable posterior characterization. Conjugacy is out of question here: even in the exchangeable context it is a property characterizing the Dirichlet process (see James et al., 2006). Nevertheless, conditional on a set of suitable latent variables, the posterior distribution of the original completely random vector (μ_1, μ_2) turns out to be again a completely random vector leading to a neat posterior characterization and viable methods for sampling.

Consider a sample of n observations $(X_i)_{i=1}^n$ from $\tilde{p}_1$ with unique values $\underline{X}_n^* = (X_1^*, \ldots, X_k^*)$ and associated multiplicities $(n_1, \ldots, n_k)$; analogously, consider m observations $(Y_j)_{j=1}^m$ from $\tilde{p}_2$ with unique values $\underline{Y}_m^* = (Y_1^*, \ldots, Y_c^*)$ and multiplicities $(m_1, \ldots, m_c)$. While it is immediate to check for ties, hyper-ties cannot be identified from the data. To this end, we define a latent random element p encoding the hyper-ties, such that $p = \{(i_l, j_l)\}_l$, where (i, j), with $1 \leq i \leq k$ and $1 \leq j \leq c$,

denotes a hyper-tie between X_i^* and Y_j^*. Moreover $(i,0)$, with $1 \leq i \leq k$, denotes that X_i^* does not form a hyper-tie with any value in $\underline{Y}_m^*$ and $(0,j)$, with $1 \leq j \leq c$, denotes that Y_j^* does not form an hyper-tie with any value in $\underline{X}_n^*$.

Therefore, if $(i,j) \in p$ with $i \neq 0$ and $j \neq 0$, it means that X_i^* and Y_j^* come from the same pair of atoms in representation (3). Instead, $(i,0) \in p$ implies that X_i^* is the only value associated to a specific pair, and similarly for Y_j^* if $(0,j) \in p$. Since we are working with unique values, it is clear that each X_i^* and Y_j^* can form at most one hyper-tie, i.e. it is associated to a unique member of p. This justifies the following formal definition.

Definition 2. We say that $p = \{(i_l, j_l)\}_l$ is a *compatible hyper-ties structure* for $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$ if, firstly, for any $1 \leq i \leq k$, there exists exactly one i_l such that $i_l = i$, thus each element of $\underline{X}_n^*$ forms at most one hyper-tie; secondly, for any $1 \leq j \leq c$, there exists exactly one j_l such that $j_l = j$, thus each element of $\underline{Y}_m^*$ forms at most one hyper-tie; lastly, for any l, if $i_l = 0$ then $j_l \neq 0$, thus at least one coordinate refers to an element of $\underline{X}_n^*$ or $\underline{Y}_m^*$.

As a simple example, suppose that $\underline{X}_n$ and $\underline{Y}_m$ contain respectively 2 and 1 unique values. Then $k = 2$, $c = 1$ and the support of p is

$$\mathcal{P} = \Big\{ \{(1,1),(2,0)\}, \{(1,0),(2,1)\}, \{(1,0),(2,0),(0,1)\} \Big\}.$$

Once the latent structure p is identified, its elements can be conveniently partitioned into the set $\Delta_p = \{(i,j) \in p \mid i \neq 0 \text{ and } j \neq 0\}$, which includes all the hyper-ties, and the sets $\Delta_p^1 = \{(i,j) \in p \mid j = 0\}$ and $\Delta_p^2 = \{(i,j) \in p \mid i = 0\}$. If X_i^* and Y_j^* form a hyper-tie, it means that (X_i^*, Y_j^*) is an actual atom in representation (3). Instead, if X_i^* does not form a hyper-tie, we have a partial knowledge of the original pair: the unknown second coordinate can be sampled from $P_{X_i^*}(\cdot)$, that is the conditional distribution given X_i^*, induced by the joint measure G_0, which will henceforth be assumed to be non–atomic. A similar argument applies if Y_j^* does not form a hyper-tie.

In order to simplify notation, we set $g_{i,j} = g_0(X_i^*, Y_j^*)$, $g_{i,0} = p_0(X_i^*)$, and $g_{0,j} = p_0(Y_j^*)$, where g_0 and p_0 are the density functions of G_0 and P_0 respectively, that we assume exist with respect to suitable dominating measures. Finally, we consider the following integrals

$$\tau_{n,m}(\underline{u}) = \int_{\mathbb{R}_+^2} e^{-u_1 s_1 - u_2 s_2} s_1^n s_2^m \, \rho(\mathrm{d}s_1, \mathrm{d}s_2), \quad \underline{u} = (u_1, u_2),$$

where often n and m will be equal to n_i and m_j, with $1 \leq i \leq k$ and $1 \leq j \leq c$. For consistency, we set $n_0 = m_0 = 0$.

The key result of the section relies on a latent structure that is identified by random variables whose

conditional distributions, given $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$, are available. Indeed, these random variables are given by p, whose probability mass function is proportional to

$$\left(\prod_{(i,j)\in p} g_{i,j}\right) \int_{\mathbb{R}_+^2} u_1^{n-1} u_2^{m-1} \prod_{(i,j)\in p} \tau_{n_i,m_j}(\underline{u})\, \mathrm{e}^{-\psi_b(\underline{u})}\, \mathrm{d}\underline{u},$$

the vector (U_1, U_2), whose density on $\mathbb{R}_+^2$ is proportional to $u_1^{n-1}u_2^{m-1}\prod_{(i,j)\in p}\tau_{n_i,m_j}(\underline{u})e^{-\psi_b(u)}$, the variables $\{Z_i^x\}_i$, whose distribution is $P_{X_i^*}(\cdot)$, for any $i = 1, \dots k$, and $\{Z_j^y\}_j$, whose distribution is $P_{Y_j^*}(\cdot)$, for any $j = 1, \dots, c$. We are now ready to state the key posterior characterization.

Theorem 1. *Let*$(X_i)_{i=1}^n$ *and* $(Y_j)_{j=1}^m$ *be from model* (12)*, with* Q *being the law of a n-FuRBI. Then, the distribution of* (μ_1, μ_2) *conditional on* $(X_i)_{i=1}^n$*,* $(Y_j)_{j=1}^m$ *and the set of latent variables* $(p, U_1, U_2,$ $\{Z_i^x\}_i, \{Z_j^y\}_j)$ *is*

$$(\hat{\mu}_1, \hat{\mu}_2) + \sum_{(i,j)\in\Delta_p} J_{i,j}\delta_{\left(X_i^*, Y_j^*\right)} + \sum_{(i,j)\in\Delta_p^1} J_{i,0}\delta_{\left(X_i^*, Z_i^x\right)} + \sum_{(i,j)\in\Delta_p^2} J_{0,j}\delta_{\left(Z_j^y, Y_j^*\right)},$$

where $(\hat{\mu}_1, \hat{\mu}_2)$ *is a completely random vector with intensity* $e^{-U_1 s_1 - U_2 s_2}\rho(\mathrm{d}s_1, \mathrm{d}s_2)G_0(\mathrm{d}x)$ *and* $J_{i,j} = (J_{i,j}^1, J_{i,j}^2)$*, with* $i = 0, \dots, k$ *e* $j = 0, \dots, c$*, are jumps with density proportional to*

$$s_1^{n_i} s_2^{m_j} e^{-U_1 s_1 - U_2 s_2}\rho(\mathrm{d}s_1, \mathrm{d}s_2).$$

Moreover $(\hat{\mu}_1, \hat{\mu}_2)$ *and* $J_{i,j}$ *are independent.*

Conditional on the latent variables, the structure is quite intuitive: the posterior is the law of a completely random vector with modified intensity and fixed locations, given by the pairs formed by the hyper-ties. This is somehow reminiscent of the posterior structures of exchangeable models (James et al., 2009; Lijoi and Prünster, 2010), with the key novelty played by the new notion of hyper-ties, in addition to the identification of a suitable latent structure.

The distribution of the latent variables admits a nice interpretation. For instance, the mass function of the latent structure p is the product of two terms: the probability of observing the number of hyper-ties identified by p times the likelihood that exactly those pairs are formed, through the density function g_0. Thus, thanks to the homogeneity of the original completely random vector, we observe a separate effect for jumps and locations on this hidden clustering structure. The next corollary shows how the posterior distribution of the normalized measures can be deduced from Theorem 1. The statement focuses on p_1, though an analogous representation holds also for p_2.

Corollary 2. *Under the same assumptions of* Theorem 1*, conditional on* $(X_i)_{i=1}^n$*,* $(Y_j)_{j=1}^m$ *and the latent variables* $(p, U_1, U_2, \{Z_i^x\}_i, \{Z_j^y\}_j)$*, the random probability measure* p_1 *in* (3) *equals in distribution*

$$w_1 \frac{\hat{\mu}_1}{T_1} + w_2 \frac{\sum_{(i,j)\in\Delta_p} J_{i,j}^1 \delta_{\left(X_i^*, Y_j^*\right)}}{\sum_{(i,j)\in\Delta_p} J_{i,j}^1} + w_3 \frac{\sum_{(i,j)\in\Delta_p^1} J_{i,0}^1 \delta_{\left(X_i^*, Z_i^x\right)}}{\sum_{(i,j)\in\Delta_p^1} J_{i,0}^1} + w_4 \frac{\sum_{(i,j)\in\Delta_p^2} J_{0,j}^1 \delta_{\left(Z_j^y, Y_j^*\right)}}{\sum_{(i,j)\in\Delta_p^2} J_{0,j}^1},$$

where $T_1 = \hat{\mu}_1(\mathbb{X}\times\mathbb{X})$*, while*

$$w_1 \propto T_1, \quad w_2 \propto \sum_{(i,j)\in\Delta_p} J_{i,j}^1, \quad w_3 \propto \sum_{(i,j)\in\Delta_p^1} J_{i,0}^1, \quad w_4 \propto \sum_{(i,j)\in\Delta_p^2} J_{0,j}^1,$$

with the constraint $\sum_{i=1}^4 w_i = 1$*.*

5.2 Predictive structure

Prediction of new observations arises naturally within the Bayesian framework, since it coincides with the estimate of the distribution under a square loss function. Moreover, it has the merit of providing intuition on how the model behaves and learns and it can be used to develop marginal algorithms that avoid the direct sampling of $\tilde{p}_1$ and $\tilde{p}_2$, which are infinite-dimensional objects. In Proposition 5 we saw how to sample the first pair of observations. The next result tackles the general case.

Theorem 2. *Consider samples* $(X_i)_{i=1}^n$ *and* $(Y_j)_{j=1}^m$ *from model* (12)*, with the same setting of Theorem 1. Then there exist probability weights* ξ_0*,* $\{\xi_i^x\}$ *and* $\{\xi_j^y\}$ *such that*

$$\mathbb{P}\big(X_{n+1}\in C \mid (X_i)_{i=1}^n,\, (Y_j)_{j=1}^m\big) = \xi_0 P_0(C) + \sum_{i=1}^k \xi_i^x \delta_{X_i^*}(C) + \sum_{j=1}^c \xi_j^y P_{Y_j^*}(C)\,.$$

Analogously, there exist probability weights η_0*,* $\{\eta_i^x\}$ *and* $\{\eta_j^y\}$ *such that for any* $C\in\mathcal{X}$

$$\mathbb{P}\big(Y_{m+1}\in C \mid (X_i)_{i=1}^n,\, (Y_j)_{j=1}^m\big) = \eta_0 P_0(C) + \sum_{j=1}^c \eta_j^y \delta_{Y_j^*}(C) + \sum_{i=1}^k \eta_i^x P_{X_i^*}(C)\,.$$

Explicit formulae for the weights are available in the proof of Theorem 2, in Section $S1$. In specific cases they can be computed in closed form, conditional to the latent variables: see e.g. example $S1$ in Section $S4$ for the Inverse Gaussian case with equal jumps.

Hence, the marginal predictive distributions have a quite intuitive form: they are linear combinations

of the centering distribution P_0, a weighted version of the empirical distribution and a last term that depends on the other sample. The crucial differences with respect to prediction rules arising in the exchangeable case (Lijoi and Prünster, 2010; De Blasi et al., 2015) is the addition of the last term, which clearly shows how posterior inference changes when incorporating heterogeneous information and performing borrowing of information.

Example 6 (n-FuRBI with equal atoms)**.** If the joint distribution G_0 is degenerate such that the atoms are completely shared between $\tilde{p}_1$ and $\tilde{p}_2$, then $P_Z(\cdot) = \delta_Z(\cdot)$. Therefore, the last term in Theorem 2 becomes a weighted version of the empirical distribution relative to the other sample.

Algorithms for posterior inference and prediction are derived in Section S4.

6 Numerical Illustrations and Real Data Analyses

6.1 Bayesian mixture models

Discrete Bayesian models, as the one specified in (12), are usually not employed directly on the data, but as a building block in hierarchical mixture models: in this setting X and Y are hidden values that describes the clustering structure within the data. Such models have been introduced by Lo (1984) for the Dirichlet processes and gained popularity thanks also to the availability of sampling methods for posterior inference (Escobar and West, 1995; Ishwaran and James, 2001; Neal, 2000). Suppose $\{f(\cdot \mid x) : x \in \mathbb{X}\}$ is a family of probability density kernels on a space $\mathbb{W}$. Then the model can be formulated in the context of (12) as

$$\begin{matrix} W_i \mid X_i \overset{\text{ind}}{\sim} f(\cdot \mid X_i) \\ X_i \mid \tilde{p}_1 \overset{\text{i.i.d.}}{\sim} \tilde{p}_1 \end{matrix}, \quad \begin{matrix} V_j \mid Y_j \overset{\text{ind}}{\sim} f(\cdot \mid Y_j) \\ Y_j \mid \tilde{p}_2 \overset{\text{i.i.d.}}{\sim} \tilde{p}_2 \end{matrix}, \quad (\tilde{p}_1, \tilde{p}_2) \sim \text{n-FuRBI}.$$

where $(W_i)_{i=1}^n$ and $(V_j)_{j=1}^m$ are the observable samples and are assumed to be conditionally independent, given $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$. Integrating out the latent variables $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$, the data are random draws from suitable countable mixtures, i.e.

$$W_i \mid \tilde{p}_1 \overset{iid}{\sim} \int f(\cdot \mid x)\, \tilde{p}_1(\mathrm{d}x), \quad V_j \mid \tilde{p}_2 \overset{iid}{\sim} \int f(\cdot \mid y)\, \tilde{p}_2(\mathrm{d}y).$$

Example 7 (Gaussian mixtures)**.** We assume $f(\cdot \mid x) := N(\cdot \mid x, \sigma^2)$, with σ^2 positive known constant, to be the normal density. Thus, the latent parameter is the mean, i.e. $\mathbb{X} = \mathbb{R}$. In this case $\text{cov}(X_i, Y_j) = \text{cov}(W_i, V_j)$, so that the joint behavior of the latent means is reflected on the

observations: this shows the importance of the correlation structure given by Proposition 4 also for hierarchical models. Alternatively, the latent parameters could specify both the mean and the variance, with $\mathbb{X} = \mathbb{R} \times \mathbb{R}_+$.

The goal is then to draw samples from the posterior distribution given $(W_i)_{i=1}^n$ and $(V_j)_{j=1}^m$: however this requires to integrate out all the possible partitions of the $n + m$ latent variables. As detailed in Section S4., it is possible to devise a Gibbs sampler for drawing from the posterior distribution of $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$.

Once a posterior sample $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$ is generated, relevant quantities of interest can be approximated by exploiting the conditional independence of $(W_i)_{i=1}^n$ and $(V_j)_{j=1}^m$, given the latent variables.

6.2 Simulation study for density estimation

We consider a simple application with simulated data, in order to understand how inference changes when taking into account heterogeneous sources of information. Assume the following generating mechanism: $W_i \overset{\text{i.i.d.}}{\sim} N(\cdot \mid 10, 1)$, for $i = 1, \dots, 20$, and $V_j \overset{\text{i.i.d.}}{\sim} N(\cdot \mid -10, 1)$, for $j = 1, \dots, 100$. Supposing only the phenomenon associated to the first sample is of interest, hierarchical mixtures are considered to make prediction on the unknown density of W_i. The kernel considered is the one specified in Example 7, with known $\sigma^2 = 1$ and latent mean μ. Four different approaches for modelling dependence between $(W_i)_{i\geq 1}$ and $(V_i)_{i\geq 1}$ are devised: the exchangeable approach, according to which sequences W and V are supposed to form one exchangeable sequence, inducing the highest positive correlation between W_i and V_j; the independent approach, according to which the sample $(V_i)_{i\geq 1}$ is disregarded entirely, that is $(W_i)_{i\geq 1}$ and $(V_i)_{i\geq 1}$ are treated independently; the hierarchical approach, where we use a hierarchical Dirichlet process (see Example 1) that corresponds to a classical borrowing of information; the FuRBI approach, where the underlying random probability measures $\tilde{p}_1$ and $\tilde{p}_2$ are n-FuRBI with equal weights and the distribution on the atoms is $G_0(\cdot \mid \rho_0) = N_2(\cdot \mid \underline{0}, 1, \rho_0)$ with $\rho_0 \sim \text{Unif}([-1, 1])$, where $N_2(\cdot \mid \underline{m}, \sigma_0^2, \rho_0)$ denotes the bivariate normal distribution with mean vector $\underline{m}$, common variance σ_0^2 and correlation ρ_0. It can be proven that under this specification $\text{corr}(W_i, V_j) = 0$, so that a priori W and V are marginally uncorrelated. The prior specification is purposely simple, especially regarding the base measure and the concentration parameter, in order to single out the effect of the borrowing between the two groups as much as possible.

For the first two cases and the n-FuRBI, the marginal distribution is given by a Dirichlet process with $\theta = 1$ and $P_0(\cdot) = N(\cdot \mid 0, 1)$; instead for the hierarchical process the concentration parameters are

fixed in order to match the expected number of different clusters with the other methods, for a fair comparison. As highlighted in Example 5, n-FuRBI with equal jumps lead to the most general setting in terms of achievable correlation between samples; moreover, choosing the marginal processes to derive from a Gamma process, we can achieve any value in the interval $(-1, 1)$, tuning appropriately the concentration parameter θ.

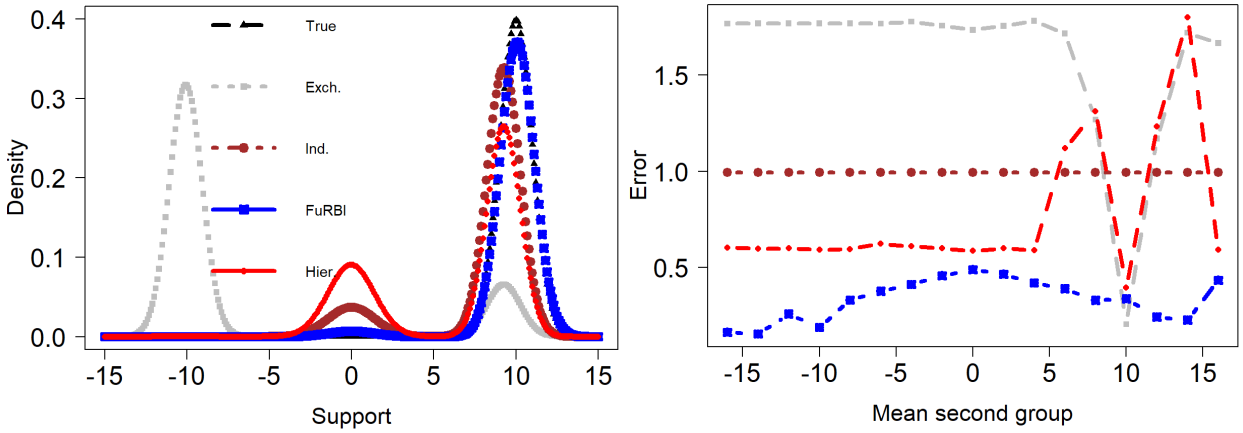


Figure 1: Left: mean posterior densities for the case with opposite true means. Right: mean integrated error (computed on a grid and as the median over 50 different samples) for the four estimates, varying the true mean of V.

The left panel of Figure 1 shows the performances of the four methods, after the application of the blocked Gibbs sampler provided in the supporting material: the mean posterior density (computed pointwise) is depicted. The exchangeable approach behaves very badly, as expected, because the two samples clearly have a different distribution. The independent choice leads to a reasonable estimate, even if it still overestimates the probability mass around the prior mean (because of the small sample size of the first sample). The hierarchical estimate is quite good, but our proposal, instead, fits almost perfectly the target density and seems to exploit the opposite behaviour of the two phenomena: this is clearly highlighted by the posterior distribution of ρ_0, whose approximated mean is close to -0.9.

One may wonder whether these superior performances follow from the precise specification above, with opposite true means. Therefore, we have repeated the experiment by keeping the same generating mechanism for W, but with the true mean of V ranging in the set $\{-16, -14, \ldots, 14, 16\}$: the mean integrated absolute error (computed on a grid and as the median over 50 different samples) is depicted in the right panel of Figure 1. It is apparent that the FuRBI approach almost always yields the smallest error, regardless of the true value. Its performance is close to the exchangeable case only when the two true means are equal, that is when exchangeability actually holds; analogously, the n-FuRBI priors yield the highest error when the mean of V corresponds to the prior mean, i.e., when the other group provides less additional information. The hierarchical process captures the right dependence when the two means coincide, but can be misled when they are close; finally,

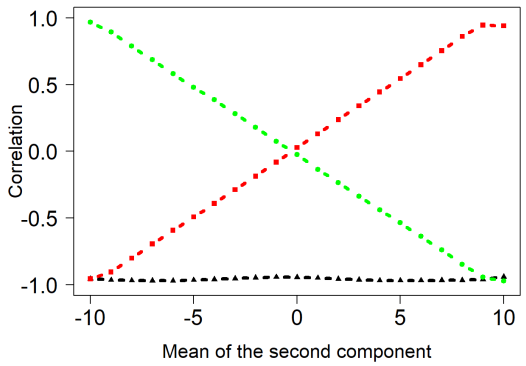


Figure 2: Posterior median of the correlation (obtained through 100 simulation studies) between the three unknown means. Black with triangular shapes: correlation between the first and second components. Red with square shapes: correlation between the first and third components. Green with circular shapes: correlation between the second and third components.

Mean of V	Exch.	Ind.	FuRBI	Hier.
-16	1.769	0.995	**0.163**	0.604
-10	1.769	0.995	**0.189**	0.592
0	1.737	0.995	**0.489**	0.587
10	**0.205**	0.995	0.338	0.397
16	1.666	0.995	**0.435**	0.592

Table 1: Mean integrated absolute error associated to the four methods for some values of the mean of V. The values in bold are the smallest ones for each row.

when the second sample is very far from the first one it performs better than the independent model, probably thanks to the different inner clustering structure. The results are also summarized in Table 1. Thus, n-FuRBI seem to be always capable of combining heterogeneous information in the right way; in particular, at least in this example, they recognize the most useful type of borrowing of information. In Section S5.1 similar experiments are conducted, using different data generating distributions: they show that the conclusions hold even when the data display significantly different features, as multimodality or heavy tails.

Finally, we consider a similar application with three groups, in order to see whether n-FuRBI are able to discern more complex types of dependence. We assume to observe $W_{1,i} \overset{\text{i.i.d.}}{\sim} N(\cdot \mid 10, 1)$, $W_{2,i} \overset{\text{i.i.d.}}{\sim} N(\cdot \mid -10, 1)$, and $W_{3,i} \overset{\text{i.i.d.}}{\sim} N(\cdot \mid x, 1)$, where $i = 1, \ldots, 20$ and $x \in \{-10, -9, \ldots, 10\}$. Then, for each value of x we apply the same n-FuRBI with the same weights described above, but where the atoms are distributed according to

$$G_0(\cdot) = N_3\left(\cdot \middle| \underline{0}, 1, \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}\right),$$

where $N_3(\cdot \mid \mu_0, \sigma^2, \Psi)$ denotes a multivariate normal distribution with mean μ_0, all the variances equal to σ^2 and correlation matrix Ψ and $\rho_{12}, \rho_{13}, \rho_{23} \overset{\text{i.i.d.}}{\sim} \text{Unif}([-1, 1])$. The posterior medians of ρ_{12}, ρ_{13} and ρ_{23} are depicted in Figure 2, for any value of x. The results are in line with our intuition:

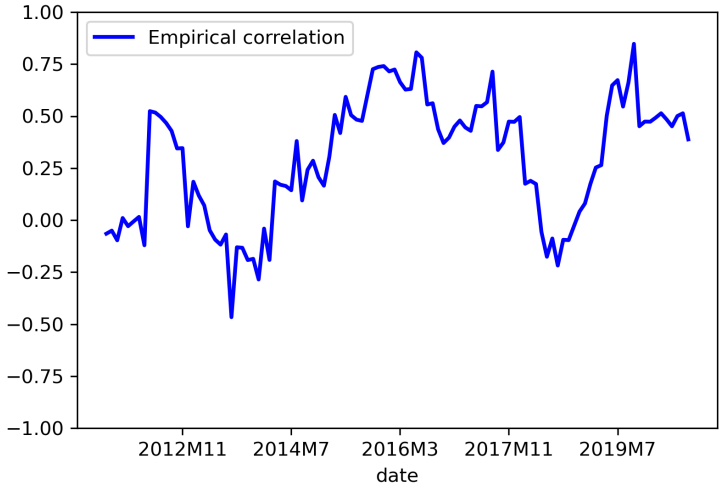


Figure 3: Empirical correlation between average stock return and average commodity return computed on a moving window of 12 months using data from March 2011 to January 2021.

the correlation between the first and second component is always close to -1 (indeed they have opposite behaviour relative to the prior), while ρ_{13} and ρ_{23} vary linearly with x, being positive when the means have the same sign.

6.3 Predicting stocks and bonds returns

Findings from the previous section and Section S5.1 suggest that n-FuRBI may be used to enhance density estimates and prediction in multi-sample data. Here, the performance is showcased on a real dataset of stock and bond returns. We collected monthly returns of January 2021 for a sample of 49 stocks portfolios from the Kenneth R. French's Data Library (data available at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`) and for a sample of 55 commodities from the Primary Commodity Prices Database of the International Monetary Fund (data available at `https://www.imf.org/en/Research/commodity-prices`).

We employ a Bayesian mixture model and assume that stock and bonds returns, denoted by W_i and V_j, respectively, are sampled from mixtures of normals where the mixing distributions act on mean and variance of the kernel, i.e.,

$$W_i \mid \tilde{p}_1 \overset{iid}{\sim} \int N(\cdot \mid x, \sigma_w^2)\, \tilde{p}_1(\mathrm{d}x, \mathrm{d}\sigma_w^2) \qquad V_j \mid \tilde{p}_2 \overset{iid}{\sim} \int N(\cdot \mid y, \sigma_v^2)\, \tilde{p}_2(\mathrm{d}y, \mathrm{d}\sigma_v^2).$$

Stocks and commodities exhibit correlation that largely varies over time ranging from positive to

negative values (see, for instance, Bhardwaj and Dunsby, 2013, and Figure 3). As a consequence, commodities returns contain useful information to make inference over the distribution of stocks portfolios, and viceversa. Thus, borrowing of information represents a natural strategy to improve inference. However, returns may differ even largely in value between the two sets of financial instruments, especially in periods of negative correlation. For instance, in our dataset, 53% of the observed stocks returns are negative, while only 16% of the bonds returns have negative sign. As such, classical nonparametric borrowing, consisting in sharing of mixture components, is not appropriate and, as shown in the following, possibly harmful. We instead make use of n-FuRBI models as prior distribution, i.e.,

$$\begin{aligned}(\tilde{p}_1, \tilde{p}_2) \mid \theta, z, G_0 &\sim \text{n-FuRBI}(\theta, \rho, G_0)\\ \theta &\sim \text{Gamma}(\alpha, \beta)\end{aligned}$$

The base measure G_0 is chosen so that marginal distributions are given by normalized CRMs with conjugate Normal-InverseGamma base measure, i.e.

$$\begin{aligned}G_0(\mathrm{d}_x, \mathrm{d}_y, \mathrm{d}\sigma_w^2, \mathrm{d}\sigma_v^2 \mid \rho_0) = N_2(\mathrm{d}x, \mathrm{d}y \mid m, \Sigma(\lambda_1, \lambda_2, \sigma_w^2, \sigma_v^2\rho_0))\\ \times \text{InvGamma}(\mathrm{d}\sigma_w^2 \mid \alpha_1, \beta_1) \times \text{InvGamma}(\mathrm{d}\sigma_v^2 \mid \alpha_2, \beta_2)\end{aligned}$$

with

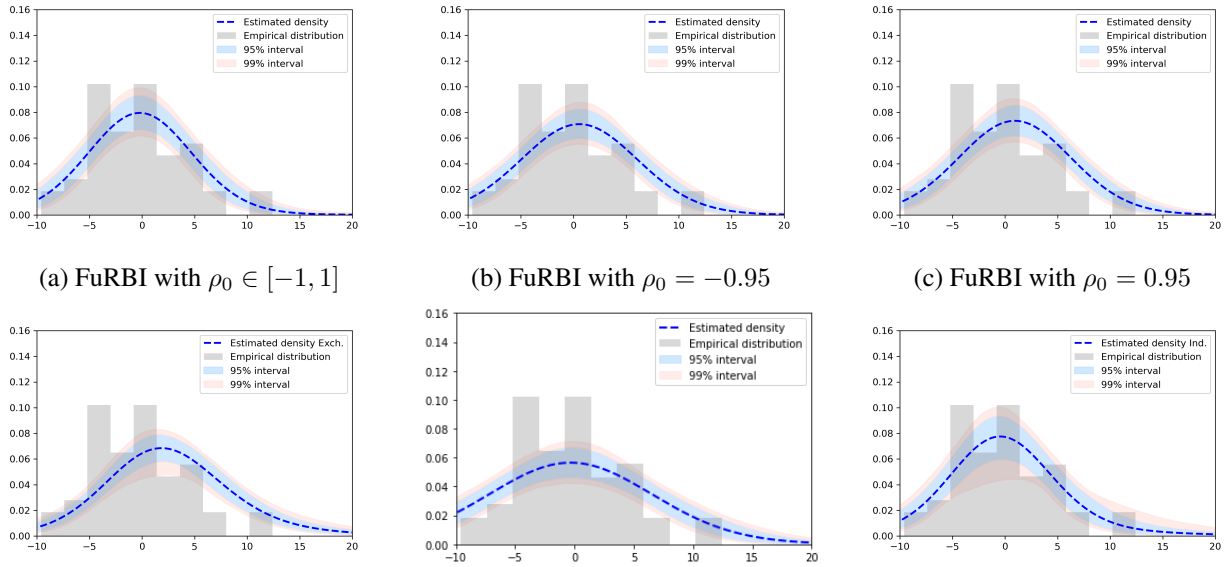
$$m = (m_1, m_2)' \qquad \text{and} \qquad \Sigma = \begin{bmatrix} \frac{\sigma_w^2}{\lambda_1} & \rho_0 \frac{\sigma_w}{\lambda_1^{1/2}} \frac{\sigma_v}{\lambda_2^{1/2}} \\ \rho_0 \frac{\sigma_w}{\lambda_1^{1/2}} \frac{\sigma_v}{\lambda_2^{1/2}} & \frac{\sigma_v^2}{\lambda_2} \end{bmatrix}$$

and we use the following joint underlying Lévy intensity $v(\mathrm{d}s_1, \mathrm{d}s_2, \mathrm{d}x_1, \mathrm{d}x_2) = \{z\,[\rho(\mathrm{d}s_1)\delta_0(\mathrm{d}s_2) + \rho(\mathrm{d}s_2)\delta_0(\mathrm{d}s_1)] + (1-z)\,\rho(\mathrm{d}s_1)\delta_{s_1}(\mathrm{d}s_2)\}\,\theta\,G_0(\mathrm{d}x_1, \mathrm{d}x_2)$, with $z \sim \text{Unif}([0,\,1])$. We term the resulting n-FuRBI *additive n-FuRBI*, since the series representation of the corresponding FuRBI CRMs is

$$\tilde{\mu}_1(\cdot) \overset{a.s.}{=} \sum_{k\geq 1} W_k \delta_{\theta_{0,k}} + \sum_{k\geq 1} J_k \delta_{\theta_{1,k}} \qquad\qquad \tilde{\mu}_2(\cdot) \overset{a.s.}{=} \sum_{k\geq 1} W_k \delta_{\phi_{0,k}} + \sum_{k\geq 1} V_k \delta_{\phi_{2,k}},$$

where $(\theta_{0,k}, \phi_{0,k}) \overset{i.i.d}{\sim} G_0$, $\theta_{1,k} \overset{i.i.d}{\sim} P_0$ and $\phi_{2,k} \overset{i.i.d}{\sim} P_0$. When G_0 is degenerate on the main diagonal (i.e. $\rho_0 = 1$), one retrieves GM-dependent completely random measures (Lijoi et al., 2014a,b; Lijoi and Nipoti, 2014). In order to obtain two Dirichlet processes marginally we set $\rho(s) = s^{-1}e^{-s}$, so that $\beta = 1/(1+\theta)$ and $\gamma = (1-z)\,{}_3F_2(\theta - \theta\,z + 2,\, 1,\, 1;\, \theta + 2,\, \theta + 2;\, 1)\theta/(1+\theta)^2$, where ${}_3F_2$ is the generalized hypergeometric function.

As for the hyperparameters of the model, we set the a priori expectations m_1 and m_2 in the two groups equal to the empirical averages of the two groups in December 2020, i.e., the month preceding the data collection, leading to $m_1 = 5.8591$ and $m_2 = 3.9731$. In the following, we say that a financial instrument is *outperforming* if its observed return is higher than its a priori expected value.



(a) FuRBI with $\rho_0 \in [-1, 1]$ (b) FuRBI with $\rho_0 = -0.95$ (c) FuRBI with $\rho_0 = 0.95$

(d) Exchangeable model (e) GM-dependent model (f) Independent model

Figure 4: Posterior density estimates for stocks returns.

In order to assign ρ_0, we use the results of Propositions 4 and 5. The elicited ρ_0 should reflect our prior opinion about the correlation, which means that it should induce a learning mechanism agreeing with the following principle: under positive/negative correlation, conditioning on the event of outperforming commodities, the prior probability of outperforming/underperforming stocks should increase. Prior opinion about the correlation can be formulated working with financial experts and, thanks to n-FuRBI, incorporated through an informative prior on the parameter ρ_0. Here, we consider three scenarios: in the first and second, we derive inferential results under a prior opinion of negative and positive correlation, respectively, while in the third scenario we assume that no information on the correlation is available. The three scenarios are obtained with, respectively, $\rho_0 = 0.95$, $\rho_0 = -0.95$, and using a uniform prior on ρ_0. After standardizing the data, we set the remaining hyperparameters in a weakly informative way, i.e. $\lambda_1 = \lambda_2 = 1$, $\alpha_1 = \alpha_2 = 2$, and $\beta_1 = \beta_2 = 4$. Sensitivity analysis, carried out in Section S6.2, shows that results are robust with respect to different choices for λ_j, α_j and β_j for $j = 1, 2$. We perform $50,000$ iterations of the marginal algorithm (Section S4.1) and discard the first $10,000$ as burn–in. Section S8. contains results about convergence diagnostic, mixing performance, and computational times of the algorithm.

Finally, we compare our approach with three alternative models: the independent model and the exchangeable model, described in the previous section, and the GM-dependent model from Lijoi et al. (2014b), which performs classical borrowing based on ties and shares the same additive structure of additive n-FuRBI.

Figure S5 displays the posterior density estimates for stocks returns. The analogous figure for bonds

	ALCPO	MLCPO
FuRBI $\rho_0 \in [-1, 1]$	**-1.2347**	**-0.9627**
FuRBI $\rho_0 = -0.95$	-1.2925	-1.0115
FuRBI $\rho_0 = 0.95$	-1.2896	-1.0149
Exch	-1.5024	-1.1521
GM-dep	-1.4864	-1.1557
Ind	-1.3495	-1.1017

Table 2: ALCPO and MLCPO under the three models. Best performance is highlighted in bold.

returns can be found in Section S6.1. Models employing additive n-FURBI produce density estimates that better resemble the empirical distribution. The best performance is attained by placing a (non-informative) prior over the correlation ρ_0, which leads to a posterior skewed towards negative values but still quite dispersed (see Figure S7) reflecting the direction and intensity of the borrowing of information. The FuRBI models with fixed ρ_0 perform worse compared to full-borrowing; nonetheless, thanks to their flexibility, they still produce better results than other competitors. The GM-dependent and the exchangeable models yield the worst density estimates in terms of resemblance of the histogram, as expected. Indeed, the type of borrowing they perform differ from the one allowed by FuRBIs (even when $\rho_0 = 0.95$), as it is based on ties, which are not appropriate for the specific problem at hand. Lastly, we note that the independent model appears to provide a reasonable density estimation, but presents significantly higher uncertainty.

While Figure S5 provides insight on the model performance, an important *caveat* is in order: a too close resemblance of the empirical distribution may indicate overfitting.

To evaluate the predictive performance, we resort to the conditional predictive ordinates (CPOs) statistics (see, e.g. Gelfand et al., 1992; Barrios et al., 2013). Essentially, for each value i, we train the model without the i-th observation and compute the predictive density at the observed point. For the first sample it reads $\text{CPO}_i^w = \tilde{f}(w_i \mid w^{-i}, v)$, for $i = 1, \ldots, n$ and analogously for the second sample we have $\text{CPO}_j^v = \tilde{f}(v_j \mid w, v^{-j})$, for $j = 1, \ldots, m$, where w and v denote the vectors of observed returns for, respectively, stocks and commodities.

Table 2 displays the average logarithmic CPO (ALCPO) and the median logarithmic CPO (MLCPO) in the overall sample. Higher values correspond to a better performance, and the n-FuRBI exhibits the best performance.

6.4 Clustering of multivariate data with missing entries

We now show how to leverage on our methodology to perform borrowing of information and clustering with multivariate data affected by missing entries. The n-FuRBI priors are very well suited for this problem: indeed, incomplete observations can be interpreted as projections of latent complete observations and, in particular, hyper-ties between incomplete observations can be thought of as actual ties between complete observations.

We consider a P-variate ($P > 1$) dataset with missing entries and divide the dataset into distinct samples based on the missing entries: denote by $(\underline{W}_i^{(j_1,\dots,j_l)}, i = 1, \dots, n_{(j_1,\dots,j_l)})$ the sample where l outcomes with labels $(j_1, \dots, j_l)$ are missing. The dimension of the vector $\underline{W}_i^{(j_1,\dots,j_l)}$ is therefore $P_{j_1,\dots,j_l} = P - l$. Denote by $\tilde{q}_{j_1,\dots,j_l}$ the corresponding unknown distribution, i.e.,

$$\underline{W}_i^{(x)} \mid \tilde{q}_x \overset{iid}{\sim} \tilde{q}_x \qquad \text{for } i = 1, \dots, n_x \text{ and } x \in I,$$

where I is the index set of all the possible combinations of missing variables identifying different samples, which are at most $2^P - 1$. Independent analyses for each sample should clearly be avoided and classical nonparametric borrowing cannot even be specified because the support spaces of different samples differ one from the other.

To perform clustering, we assume that each $\tilde{q}_x$ is a mixture of multivariate normal kernels with

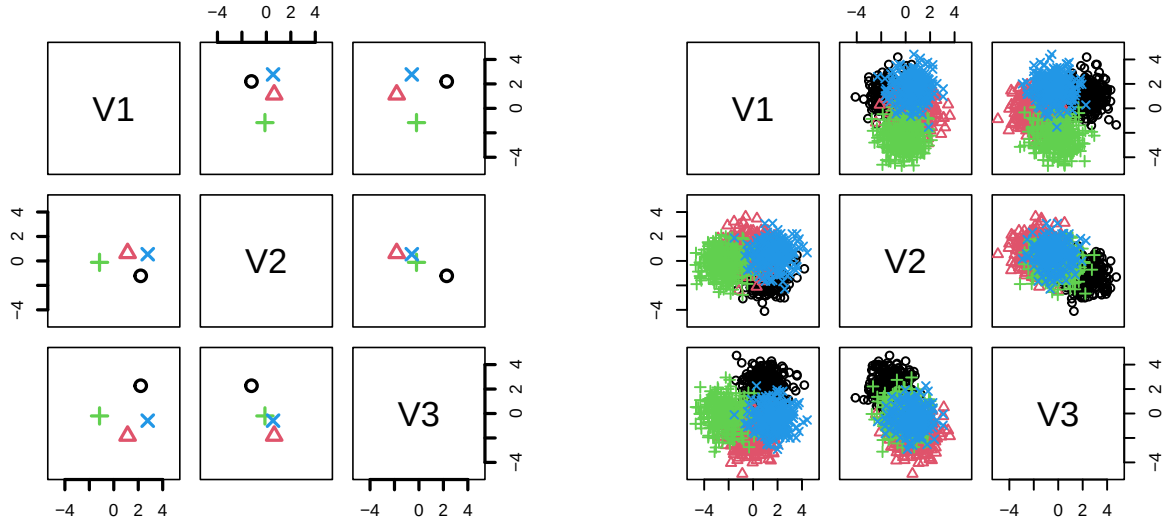


Figure 5: Simulated data: left panel shows true clusters locations, right panel shows complete simulated data for $n = 1000$ before applying the missingness mechanisms.

simul number	missing mechanism	% of missing entries	n-FuRBI $z = 0.2$	n-FuRBI $z = 0.5$	n-FuRBI $z = 0.8$	mice + k-means	mice + DPM
n.1	MCAR	16.1%	**0.7883**	0.7882	0.7881	0.7408	0.7734
n.2	MNAR	16.7%	0.7703	0.7704	**0.7706**	0.6323	0.7617
n.3	MCAR	35.9%	**0.7292**	0.7285	0.7283	0.6786	0.7165
n.4	MNAR	34%	0.7304	0.7301	**0.7432**	0.6391	0.7328

Table 3: Rand indexes for 5 competing methods: 3 n-FuRBI models with varying parameter z, mice+k-means and mice+DPM. For n-FuRBI and mice+DPM the posterior expected value is computed averaging over the Rand indexes of all clustering configurations visited by the MCMC chain after burn-in.

simul number	missing mechanism	% of missing entries	n-FuRBI $z = 0.2$	n-FuRBI $z = 0.5$	n-FuRBI $z = 0.8$	mice + k-means	mice + DPM
n.1	MCAR	16.1%	4.24	**4.19**	4.22	3	5.48
n.2	MNAR	16.7%	**4.59**	3.29	3.37	2	5.36
n.3	MCAR	35.9%	4.38	**4.18**	4.20	3	7.01
n.4	MNAR	34.0%	4.28	**4.17**	4.59	2	5.85

Table 4: Estimated number of clusters for 5 competing methods. The posterior mean is used for n-FuRBI and mice+DPM, while the number of clusters is selected by maximizing the average silhouette for mice+k-means. The true number of clusters is equal to 4.

diagonal covariance matrix and mixing measure $\tilde{p}_x$ on locations, i.e.

$$\underline{W}_i^{(x)} \mid \tilde{p}_x, \underline{\sigma}^2 \overset{iid}{\sim} \int N_{P_x}(\cdot \mid \underline{\mu}_x, \underline{\sigma}_x^2)\, \tilde{p}_x(\mathrm{d}\underline{\mu}_x),$$

where $\underline{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_P^2)$, $\underline{\sigma}_x^2$ is the restriction of $\underline{\sigma}^2$ to all the elements besides x and $N_K(\cdot \mid \underline{\mu}, \underline{\tau}^2)$ denotes the K-variate normal distribution with mean vector $\underline{\mu}$ and diagonal covariance matrix given by $\underline{\tau}^2$. Independence of the kernel (implied by the diagonal covariance matrix) is a common assumption in clustering models for multivariate responses (see, for instance, Gao et al., 2020; Franzolini et al., 2023): in this way we are forcing the clustering structure to encode all the dependence across responses. The $\tilde{p}_x$ are distributed as

$$(\tilde{p}_x, x \in I) \sim \text{additive n-FuRBI},$$

described in Section 6.3. The atoms of $(\tilde{p}_x, x \in I)$ are costrained so that an hyper-tie can be interpreted as an actual tie between complete observations: moreover the choice of dependent weights allows to recover group-specific features, if the missingness mechanism is informative. Section S7.1 provides a discussion of this and contains the details about the choice of the hyperparameters.

First, we conduct a simulation study where data for $n = 1,000$ items, $P = 3$ responses, and $K = 4$ clusters are simulated from a mixture of Gaussian distributions. Figure 5 shows the locations of the

true clusters and the complete simulated data before deleting entries. Then, different missingness mechanisms are applied to determine the entries to be treated as missing. Missing completely at random (MCAR) scenarios are obtained by sampling missing entries uniformly, while, in missing non at random (MNAR) scenarios the probability of being missing depends on the true cluster allocation. Different combinations of missing variables define different samples: the number of samples ranges from 3 to 6 among simulation scenarios. The detailed distributions of missing values are provided in Section S7.2. Different values of the hyperparameter z of the Lévy intensity are considered. Our results are compared with those obtained with two alternative approaches, called "mice + k-means" and "mice + DPM", which follow a two-steps procedure: first one imputes missing data by chained equations as implemented in the `R` package `mice` (van Buuren and Groothuis-Oudshoorn, 2011), then, the clustering structure is estimated with, respectively, k-means and a Dirichlet process mixture. Note that the number of clusters for k-means is chosen to maximize the average silhouette. For each run of the n-FuRBI model, we perform $25,000$ iterations of the MCMC chain and discard the first half as burn-in. Section S8. contains results about convergence diagnostics, mixing performance, and computational times of the algorithm. Tables 3 and 5 summarize the performance of the models. The n-FuRBI priors outperform the alternatives in all scenarios considered, in term of estimating both the number of clusters and the clustering configuration, measured by Rand indexes between the estimated configuration and the true clustering structure. Moreover, the posterior distribution of n-FuRBI models reflects uncertainty both about the estimated clustering configuration and about the imputation mechanism, which is instead ignored by two-step procedures.

Finally, we apply the model also on the `brandsma` dataset (Snijders and Bosker (2012)), which refers to grade 8 students (age about 11 years) in elementary schools in the Netherlands (see, Brandsma and Knuver, 1989). The goal is to cluster $n = 4,106$ pupils, based on their IQ verbal score (IQV), IQ performance score (IQP), language score (LRP), and arithmetic score (APR). The number of subjects presenting missing entries is 339 out of $4,106$ (i.e., 8.26%). As before, different combinations of missing variables define different samples: the number of samples is 7 in the `brandsma` dataset. In this real data analysis, the final clustering configuration provides a lower-dimensional description of the data rather than an estimate of ideal true clusters. Data are standardized before running the model, so that the sample means and variances are equal to 0 and 1. Figure 6 shows the estimated clustering configuration obtained minimizing the variation of information loss with respect to the posterior distribution. The model identifies three clusters, which show as major tendency that groups of students performing above/below average for one of the four scores tends to perform above/below average also for the other scores. In particular, a first cluster includes 53% of the subjects, which have lower performances: indeed cluster averages of the standardized scores are IQV$= -0.371$, IQP$= -0.398$, LRP$= -0.387$, and APR$= -0.435$. Instead, the second cluster, including 44% of the subjects, retains the best students: the cluster averages of the standard-

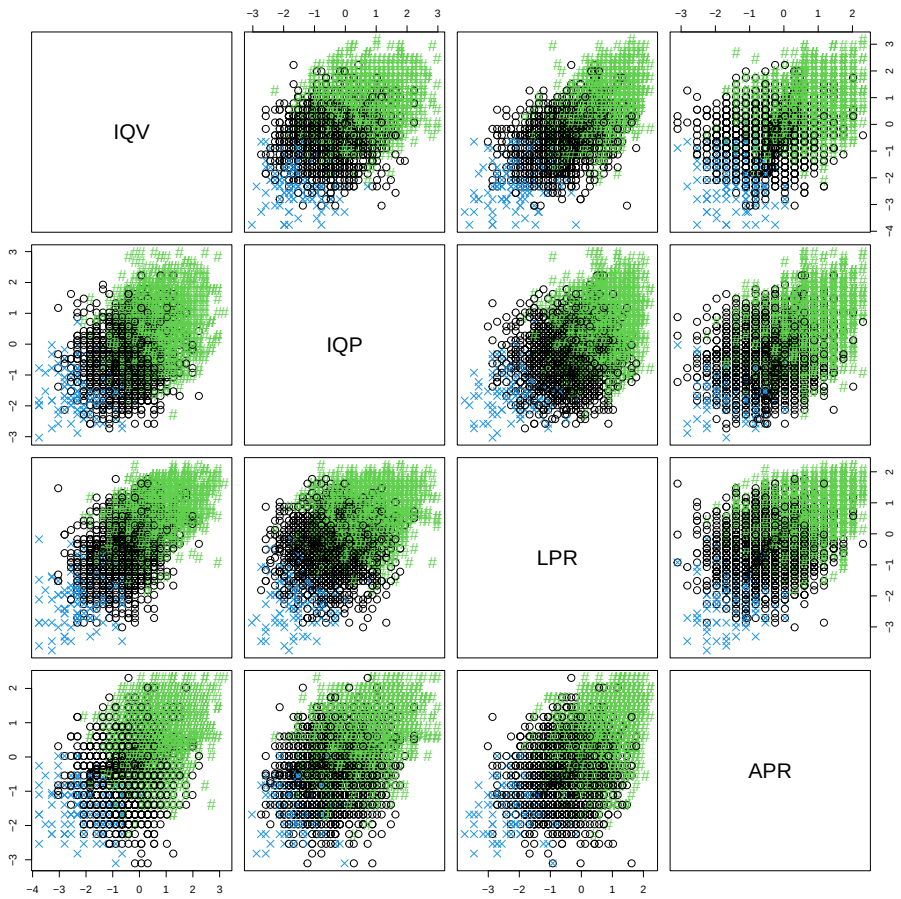


Figure 6: Scatter plots of the four scores (after standardization) for the `brandsma` dataset. Coordinates of missing data are set equal to their respective posterior median. Different colors and symbols denote the three estimated clusters obtained minimizing the variation of information loss with respect to the posterior distribution.

ized scores are IQV$= 0.609$, IQP$= 0.595$, LRP$= 0.629$, and APR$= 0.642$. Finally, the students with the worst scores are allocated to a third cluster whose averages are IQV$= -2.01$, IQP$= -1.43$, LRP$= -1.90$, and APR$= -1.34$.

7 Conclusion

Hyper-ties play a crucial role in driving the Bayesian learning mechanism and the borrowing of information across samples. However existing nonparametric priors either do not allow an explicit evaluation of the probability of a hyper-tie or, when they do, often only non-negative correlation is induced. On the contrary, n-FuRBIs allow for analytical tractability and may induce either positive or negative correlation between the random probabilities as well as across samples resulting in a

novel and flexible idea of borrowing of strength. They are immediately applicable to model multi-sample data through mixture models, as shown in Section 6.3. Morever, n-FuRBIs also allow for a variety of interesting extensions, since they can be seen as an effective building block to model non-trivial dependencies in more complex data analyses. Future work will further explore these applications.

Acknowledgements

F. Ascolani, A. Lijoi and I. Prünster are partially supported by MIUR, PRIN Project P2022H5WZ9. Part of this work was carried out while B. Franzolini was a Research Fellow at the Agency for Science, Technology and Research, in Singapore, Republic of Singapore. B. Franzolini is supported by PNRR - PE1 FAIR - CUP B43C22000800006.

References

Arbel, J. and I. Prünster (2017). A moment-matching Ferguson & Klass algorithm. *Statistics and Computing 27*(1), 3–17.

Barrios, E., A. Lijoi, L. E. Nieto-Barajas, and I. Prünster (2013). Modeling with normalized random measure mixture models. *Statistical Science 28*(3), 313–334.

Bhardwaj, G. and A. Dunsby (2013). The business cycle and the correlation between stocks and commodities. *Journal of Investment Consulting 14*(2), 14–25.

Brandsma, H. and J. Knuver (1989). Effects of school and classroom characteristics on pupil progress in language and arithmetic. *International Journal of Educational Research 13*(7), 777–788.

Brillinger, D. R. (2002). John W. Tukey: his life and professional contributions. *The Annals of Statistics 30*(6), 1535–1575.

Camerlenghi, F., D. B. Dunson, A. Lijoi, I. Prünster, and A. Rodriguez (2019). Latent nested non-parametric priors. *Bayesian Analysis 14*(4), 1303–1356.

Camerlenghi, F., A. Lijoi, P. Orbanz, and I. Prünster (2019). Distribution theory for hierarchical processes. *The Annals of Statistics 47*(1), 67–92.

Camerlenghi, F., A. Lijoi, and I. Prünster (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistics 45*(4), 1062–1091.

Catalano, M., H. Lavenant, A. Lijoi, and I. Prünster (2023). A Wasserstein index of dependence for random measures. *Journal of the American Statistical Association*, forthcoming.

Catalano, M., A. Lijoi, and I. Prünster (2021). Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *The Annals of Statistics 49*(5), 2916–2947.

Cifarelli, D. M. and E. Regazzini (1978). Nonparametric statistical problems under partial exchangeability: The role of associative means. *Quaderni Istituto Matematica Finanziaria dell'Università di Torino Serie III 12*, 1–36.

De Blasi, P., S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis & Machine Intelligence 37*(2), 212–229.

de Finetti, B. (1938). Sur la condition d'equivalence partielle. *Actualités Scientifiques et Industrielles 739*, 5–18, Translated In: Studies in Inductive and Probability, II. Jeffrey, R. (ed.) University of California Press: Berkeley 1980.

Dunson, D. and J. Park (2008). Kernel stick-breaking processes. *Biometrika 95*(2), 307–323.

Efron, B. and C. Morris (1977). Stein's paradox in statistics. *Scientific American 236*(5), 119–127.

Epifani, I. and A. Lijoi (2010). Nonparametric priors for vectors of survival functions. *Statistica Sinica 20*(4), 1455–1484.

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

Favaro, S., A. Lijoi, C. Nava, B. Nipoti, I. Pruenster, and Y. W. Teh (2016). On the stick-breaking representation for homogeneous NRMIs. *Bayesian Analysis 11*(3), 697–724.

Ferguson, T. S. and M. J. Klass (1972). A representation of independent increment processes without gaussian components. *The Annals of Mathematical Statistics 43*(5), 1634–1643.

Foti, N. J. and S. A. Williamson (2013). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE Transactions on Pattern Analysis & Machine Intelligence 37*(2), 359–371.

Franzolini, B., A. Cremaschi, W. v. d. Boom, and M. De Iorio (2023). Bayesian clustering of multiple zero-inflated outcomes. *Philosophical Transactions of the Royal Society A 381*, 1–16.

Gao, L. L., J. Bien, and D. Witten (2020). Are clusterings of multiple data views independent? *Biostatistics 21*(4), 692–708.

Gelfand, A. E., D. K. Dey, and H. Chang (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Technical Report*, Department of Statistics, Stanford University.

Gong, M., P. Liu, F. C. Sciurba, P. Stojanov, D. Tao, G. C. Tseng, K. Zhang, and K. Batmanghelich (2021). Unpaired data empowers association tests. *Bioinformatics 37*(6), 785–792.

Griffin, J. E., M. Kolossiatis, and M. F. Steel (2013). Comparing distributions by using dependent normalized random-measure mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(3), 499–529.

Griffin, J. E. and F. Leisen (2017). Compound random measures and their use in Bayesian nonparametrics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*, 525–545.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*(453), 161–173.

James, L. F., A. Lijoi, and I. Prünster (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scandinavian Journal of Statistics 33*(1), 105–120.

James, L. F., A. Lijoi, and I. Prünster (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics 36*(1), 76–97.

James, L. F., A. Lijoi, and I. Prünster (2010). On the posterior distribution of classes of random means. *Bernoulli 16*(1), 155–180.

Kingman, J. (1967). Completely random measures. *Pacific Journal of Mathematics 21*(1), 59–78.

Kingman, J. (1993). *Poisson Processes*. Clarendon Press, Oxford.

Lee, A. M., B.-E. Sæther, and S. Engen (2020). Spatial covariation of competing species in a fluctuating environment. *Ecology 101*(1), e02901.

Lijoi, A., R. H. Mena, and I. Prünster (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association 100*(472), 1278–1291.

Lijoi, A. and B. Nipoti (2014). A class of hazard rate mixtures for combining survival data from different experiments. *Journal of the American Statistical Association 109*(506), 802–814.

Lijoi, A., B. Nipoti, and I. Prünster (2014a). Bayesian inference with dependent normalized completely random measures. *Bernoulli 20*(3), 1260–1291.

Lijoi, A., B. Nipoti, and I. Prünster (2014b). Dependent mixture models: clustering and borrowing information. *Computational Statistics & Data Analysis 71*, 417–433.

Lijoi, A. and I. Prünster (2010). Models beyond the Dirichlet process. In *Bayesian nonparametrics (Hjort, N.L., Holmes, C.C., Müller, P., Walker, S.G. Eds.)*, pp. 80–136. Cambridge University Press, Cambridge.

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

Lorenz, D. J., S. Levy, and S. Datta (2018). Inferring marginal association with paired and unpaired clustered data. *Statistical methods in medical research 27*(6), 1806–1817.

MacEachern, S. N. (1999). Dependent nonparametric processes. *In ASA Proceedings of the Section on Bayesian Statistical Science.*, Alexandria, VA: American Statistical Association.

MacEachern, S. N. (2000). Dependent Dirichlet processes. *Technical Report*, The Ohio State University.

Müller, P., F. Quintana, and G. Rosner (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(3), 735–749.

Müller, P., F. A. Quintana, A. Jara, and T. Hanson (2015). *Bayesian nonparametric data analysis*. Springer.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*(2), 249–265.

Papaspiliopoulos, O. and G. O. Roberts (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika 95*(1), 169–186.

Petralia, F., V. Rao, and D. B. Dunson (2012). Repulsive mixtures. In *Advances in Neural Information Processing Systems - NIPS*.

Quinlan, J. J., F. A. Quintana, and G. L. Page (2017). Parsimonious hierarchical modeling using repulsive distributions. *arXiv preprint arXiv:1701.04457*.

Quintana, F., P. Müller, A. Jara, and S. MacEachern (2022). The dependent Dirichlet process and related models. *Statistical Science 37*(1), 24–41.

Regazzini, E., A. Lijoi, and I. Prünster (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics 31*(2), 560–585.

Rigon, T. and D. Durante (2021). Logit stick-breaking priors for bayesian density regression. *Journal of Statistical Planning and inference 211*, 131–142.

Riva-Palacio, A. and F. Leisen (2021). Compound vectors of subordinators and their associated positive Lévy copulas. *Journal of Multivariate Analysis 183*, 104728.

Rodriguez, A. and D. Dunson (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis 6*(1), 145–178.

Rodriguez, A., D. B. Dunson, and A. E. Gelfand (2008). The nested Dirichlet process. *Journal of the American Statistical Association 103*(483), 1131–1154.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica sinica 4*(2), 639–650.

Snijders, T. and R. Bosker (2012). Multilevel analysis. *Netherlands: SAGE Publications*.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association 101*(476), 1566–1581.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software 45*(3), 1–67.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Comm. Statist. Simulation Comput. 36*(1-3), 45–54.

Appendix

S1 Proofs

S1.1 Proofs of Section 1

Proof of Proposition 1. Consider two partially exchangeable sequences X and Y whose elements take value in $\mathbb{R}$. By de Finetti's representation theorem, there exist two random probability measures $\tilde{p}_1$ and $\tilde{p}_2$ such that

$$(X_i, Y_j) \mid \tilde{p}_1, \tilde{p}_2 \overset{\text{iid}}{\sim} \tilde{p}_1 \times \tilde{p}_2.$$

Note that $\text{cov}(X_i, Y_j) = \mathbb{E}\{\text{cov}(X_i, Y_j \mid \tilde{p}_1, \tilde{p}_2)\} + \text{cov}\{\mathbb{E}(X_i \mid \tilde{p}_1), \mathbb{E}(Y_j \mid \tilde{p}_2)\}$, where the first term equals 0, so that

$$\text{cov}(X_i, Y_j) = \text{cov}\left(\int x\, \tilde{p}_1(dx), \int x\, \tilde{p}_2(dx)\right),$$

and analogously

$$\text{cov}(X_i, X_{i'}) = \text{cov}\left(\int x\, \tilde{p}_1(dx), \int x\, \tilde{p}_1(dx)\right) = \text{var}\left(\int x\, \tilde{p}_1(dx)\right).$$

Lastly assume that $\tilde{p}_1 \overset{d}{=} \tilde{p}_2$, where $\overset{d}{=}$ indicates equality in distribution. By the Cauchy-Schwartz inequality

$$-\text{var}\left(\int x\, \tilde{p}_1(dx)\right) \leq \text{cov}\left(\int x\, \tilde{p}_1(dx), \int x\, \tilde{p}_2(dx)\right) \leq \text{var}\left(\int x\, \tilde{p}_1(dx)\right),$$

which, in terms of the observables, can be equivalently rewritten as

$$-\text{cov}(X_i, X_{i'}) \leq \text{cov}(X_i, Y_j) \leq \text{cov}(X_i, X_{i'}).$$

□

Proof of Proposition 2. By definition of covariance we have

$$\text{cov}(X_i, Y_j) = \text{cov}\left(\sum_{j\geq 1} J_j \theta_j, \sum_{k\geq 1} W_k \phi_k\right) = \sum_{j\geq 1}\sum_{k\geq 1} \text{cov}\left(J_j\theta_j, W_k\phi_k\right).$$

For arbitrary j and k we have

$$E\left(J_j W_k \theta_j \phi_k\right) = E(J_j W_k) E(\theta_j \phi_k) \geq E(J_j W_k) E(\theta_j) E(\phi_k),$$

since $\text{cov}(\theta_j, \phi_k) \geq 0$. Denoting $c = E(\theta_j) = E(\phi_k)$, we get

$$\text{cov}\left(J_j\theta_j, W_k\phi_k\right) \geq c^2\text{cov}(J_j, W_k).$$

Finally, since $\tilde{p}_1$ and $\tilde{p}_2$ are random probability measures it holds

$$\text{cov}(X_i, Y_j) \geq c^2\text{cov}\left(\sum_{j\geq 1} J_j, \sum_{k\geq 1} W_k\right) = 0,$$

which completes the proof. □

S1.2 Proofs of Section 2

Proof of Proposition 3. Recall that

$$\beta := \sum_{k\geq 1} \text{E}(\bar{J}_k^2) = \sum_{k\geq 1} \text{E}(\bar{W}_k^2) \qquad \gamma := \sum_{k\geq 1} \text{E}(\bar{J}_k\bar{W}_k).$$

Since

$$\text{E}(\bar{J}_k\bar{W}_k) \leq \sqrt{\text{E}(\bar{J}_k^2)\text{E}(\bar{W}_k^2)} = \text{E}(\bar{J}_k^2)$$

it follows that $\gamma \leq \beta$. Moreover, the equality holds if and only if $\bar{J}_k \overset{a.s}{=} a_k + \bar{W}_k$, for any k, with $a_k \in \mathbb{R}$. However the equality of marginal distributions implies $a_k = 0$. □

of Proposition 4. Recall that

$$\text{cov}(X_i, Y_j) = \text{cov}\left(\sum_{k\geq 1} \bar{J}_k\theta_k, \sum_{h\geq 1} \bar{W}_h\phi_h\right) = \sum_{k\geq 1}\sum_{h\geq 1} \text{cov}\left(\bar{J}_k\theta_k, \bar{W}_h\phi_h\right).$$

and for arbitrary k and h, we have

$$\begin{aligned}\mathbb{E}(\bar{J}_k\bar{W}_h\theta_k\phi_h) =& \text{E}(\bar{J}_k\bar{W}_h)\text{E}(\theta_k\phi_h) \\ =& \text{E}(\bar{J}_k\bar{W}_h)\left[\text{E}(\theta_k\phi_k)\mathbb{1}_{\{k=h\}} + \text{E}(\theta_k)\text{E}(\phi_h)\mathbb{1}_{\{k\neq h\}}\right],\end{aligned}$$

while

$$\mathbb{E}(\bar{J}_k\theta_k) = \text{E}(\bar{J}_k)\text{E}(\theta_k).$$

Thus, setting $c = E(\theta_k) = E(\phi_h)$, we have

$$\text{cov}(X_i, Y_j) = \sum_{k\geq 1} \text{E}(\bar{J}_k \bar{W}_h)\text{E}(\theta_k \phi_k) - c^2 \sum_{k\geq 1} \text{E}(\bar{J}_k)\text{E}(\bar{W}_k) + +c^2 \sum_{k\geq 1}\sum_{h\neq k} \text{cov}\left(\bar{J}_k, \bar{W}_h\right)$$

where

$$\begin{aligned}\sum_{k\geq 1}\sum_{h\neq k} \text{cov}\left(\bar{J}_k, \bar{W}_h\right) =& \text{cov}\left(\sum_{k\geq 1} \bar{J}_k, \sum_{h\geq 1} \bar{W}_h\right) - \sum_{k\geq 1} \text{cov}\left(\bar{J}_k \bar{W}_k\right) \\ =& -\sum_{k\geq 1} \text{E}(\bar{J}_k \bar{W}_h) + \sum_{k\geq 1} \text{E}(\bar{J}_k)\text{E}(\bar{W}_k)\end{aligned}$$

Putting everything together we obtain

$$\text{cov}(X_i, Y_j) = \sum_{k\geq 1} \text{E}[\bar{J}_k \bar{W}_k]\text{cov}(\theta_k, \phi_k).$$

Moreover

$$\text{var}(X_i) = \text{var}(Y_j) = \int\int x\, G_0(\text{d}x, \text{d}y) = \text{var}(\theta_k)$$

Thus, $\text{corr}(X_i, Y_j) = \gamma\, \rho_0$ proving the second statement in Proposition 4. Finally, applying the same procedure marginally, we get

$$\text{cov}(X_i, X_i') = \sum_{k\geq 1} \text{E}(\bar{J}_k^2)\, \text{var}(\theta_k)$$

which proves the first statement in Proposition 4. □

Proof of Corollary 1. The result immediately follows from Propositions 3 and 4. □

Proof of Proposition 5. Let β be the probability of a tie. By definition we get

$$\begin{aligned}\mathbb{P}\left(X_1 \in A, X_2 \in B\right) =& \mathbb{P}(X_1 \in A, X_2 \in B \mid X_1 = X_2)\beta + \\ &+ \mathbb{P}(X_1 \in A, X_2 \in B \mid X_1 \neq X_2)(1-\beta),\end{aligned}$$

which, by independence of the atoms, equals

$$\begin{aligned}\mathbb{P}\left(X_1 \in A, X_2 \in B\right) =& \mathbb{P}(X_1 \in A \in B)\beta + \\ &+ \mathbb{P}(X_1 \in A)\mathbb{P}(X_2 \in B)(1-\beta).\end{aligned}$$

Analogously, we have

$$\begin{aligned}\mathbb{P}\left(X_1 \in A, Y_1 \in B\right) =& \mathbb{P}(X_1 \in A, Y_1 \in B \mid X_1 \text{ and } Y_1 \text{ form an hyper-tie })\gamma + \\ &+ \mathbb{P}(X_1 \in A, Y_1 \in B \mid X_1 \text{ and } Y_1 \text{ do not form an hyper-tie })(1-\gamma),\end{aligned}$$

where γ is the probability of a hyper-tie, which equals

$$\begin{aligned}\mathbb{P}\left(X_1 \in A, Y_1 \in B\right) =& \mathbb{P}((X_1, Y_1) \in A \times B \mid X_1 \text{ and } Y_1 \text{ form an hyper-tie })\gamma + \\ &+ \mathbb{P}(X_1 \in A)\mathbb{P}(Y_1 \in B)(1-\gamma).\end{aligned}$$

□

S1.3 Proofs of Section 4

Proof of Proposition 6. The first point follows from the Lévy-Khintchine representation of the Laplace functional of a CRV. As for (ii), one has

$$\begin{aligned}\mathrm{E}\left(\exp\{-\lambda_1\tilde{\mu}_1(A) - \lambda_2\tilde{\mu}_2(B)\}\right) =& \mathrm{E}\left(\exp\{-\lambda_1\mu_1(A\times\mathbb{X}) - \lambda_2\mu_2(\mathbb{X}\times B)\}\right) \\ =& \mathrm{E}\big(\exp\{-\lambda_1\mu_1(A\times B^c) - \lambda_1\mu_1(A\times B) + \\ &- \lambda_2\mu_2(A^c\times B) - \lambda_2\mu_2(A\times B)\}\big).\end{aligned}$$

By independence of evaluations on disjoint sets, $\mu_1(C)$ and $\mu_2(D)$ are independent if $C \cap D = \emptyset$, so that the right hand side reads

$$\begin{aligned}\mathrm{E}\big(\exp\{-\lambda_1\tilde{\mu}_1(A) - \lambda_2\tilde{\mu}_2(B)\}\big) =& \mathrm{E}\left(\exp\{-\lambda_1\mu_1(A\times B^c)\}\right)\mathrm{E}\left(\exp\{-\lambda_2\mu_2(A^c\times B)\}\right)\times \\ &\times \mathrm{E}\left(\exp\{-\lambda_1\mu_1(A\times B) - \lambda_2\mu_2(A\times B)\}\right).\end{aligned}$$

The result follows upon upon using the expressions of the marginal and joint Laplace exponents of μ_1 and μ_2. Since from the joint Lévy intensity it is possible to recover the joint Laplace exponent, (iii) is also proved. □

In order to prove Proposition 7, we show that

$$\mathbb{P}\left(X \in A, Y \in B\right) = P_0(A)P_0(B)\left(1-\delta\right) + G_0(A\times B)\delta,$$

where

$$\delta := -\int_{\mathbb{R}_+^2}\left\{\frac{\partial^2}{\partial u_1\partial u_2}\psi_b(u_1,u_2)\right\}e^{-\psi_b(u_1,u_2)}\,\mathrm{d}u_1\mathrm{d}u_2.$$

is the probability of a pseudo-tie. We start with three Lemmas.

Lemma 1. *If ψ_b is the joint Laplace exponent of a CRV, then*

$$\int_{\mathbb{R}_+^2}\left\{\frac{\partial}{\partial u_1}\psi_b(u_1,u_2)\right\}\left\{\frac{\partial}{\partial u_2}\psi_b(u_1,u_2)\right\}e^{-\psi_b(u_1,u_2)}\,\mathrm{d}u_1\mathrm{d}u_2 = 1-\delta.$$

Proof of Lemma S2.1. Integrating by parts

$$\begin{aligned}
&\int_0^\infty \left\{\frac{\partial}{\partial u_1}\psi_b(u_1,u_2)\right\}\left\{\frac{\partial}{\partial u_2}\psi_b(u_1,u_2)\right\} e^{-\psi_b(u_1,u_2)}\,\mathrm{d}u_1 \\
&\quad = -\int_0^\infty \left\{\frac{\partial}{\partial u_2}\psi_b(u_1,u_2)\right\}\left\{\frac{\partial}{\partial u_1}e^{-\psi_b(u_1,u_2)}\right\}\mathrm{d}u_1 \\
&\quad = \left[\left[-\left\{\frac{\partial}{\partial u_2}\psi_b(u_1,u_2)\right\}e^{-\psi_b(u_1,u_2)}\right]_0^\infty + \int_0^\infty \left\{\frac{\partial^2}{\partial u_1\partial u_2}\psi_b(u_1,u_2)\right\}e^{-\psi_b(u_1,u_2)}\,\mathrm{d}u_1\right] \\
&\quad = \left[\left\{\frac{\partial}{\partial u_2}\psi_b(0,u_2)\right\}e^{-\psi_b(0,u_2)} + \int_0^\infty \left\{\frac{\partial^2}{\partial u_1\partial u_2}\psi_b(u_1,u_2)\right\}e^{-\psi_b(u_1,u_2)}\,\mathrm{d}u_1\right].
\end{aligned}$$

Note that $\int_0^\infty \left\{\frac{\mathrm{d}}{\mathrm{d}u_2}\psi_b(0,u_2)\right\}e^{-\psi_b(0,u_2)}\,\mathrm{d}u_2 = 1$, by the fundamental theorem of calculus. Thus the result follows immediately. □

Lemma 2. *We have*

$$\int_{\mathbb{R}_+^2} E\left(e^{-u_1\mu_1(\mathbb{X}\times\mathbb{X})-u_2\mu_2(\mathbb{X}\times\mathbb{X})}\mu_1(C)\mu_2(C)\right)\,\mathrm{d}u_1\mathrm{d}u_2 = G_0(C)^2\,(1-\delta) + G_0(C)\delta.$$

Proof of Lemma S2.2. By independence of evaluations on disjoint sets it follows that

$$\begin{aligned}
&\int_{\mathbb{R}_+^2} \mathrm{E}\left(e^{-u_1\mu_1(\mathbb{X}\times\mathbb{X})-u_2\mu_2(\mathbb{X}\times\mathbb{X})}\mu_1(C)\mu_2(C)\right)\,\mathrm{d}u_1\mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \mathrm{E}\left(e^{-u_1\mu_1(C)-u_2\mu_2(C)-u_1\mu_1(C^c)-u_2\mu_2(C^c)}\mu_1(C)\mu_2(C)\right)\,\mathrm{d}u_1\mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \mathrm{E}\left(e^{-u_1\mu_1(C)-u_2\mu_2(C)}\mu_1(C)\mu_2(C)\right\}\mathrm{E}\left(e^{-u_1\mu_1(C^c)-u_2\mu_2(C^c)}\right)\,\mathrm{d}u_1\mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \mathrm{E}\left(\frac{\partial}{\partial u_1}\frac{\partial}{\partial u_2}e^{-u_1\mu_1(C)-u_2\mu_2(C)}\right)\mathrm{E}\left(e^{-u_1\mu_1(C^c)-u_2\mu_2(C^c)}\right)\,\mathrm{d}u_1\mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \frac{\partial}{\partial u_1}\frac{\partial}{\partial u_2}\left[\mathrm{E}\left(e^{-u_1\mu_1(C)-u_2\mu_2(C)}\right)\right]\mathrm{E}\left(e^{-u_1\mu_1(C^c)-u_2\mu_2(C^c)}\right)\,\mathrm{d}u_1\mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \frac{\partial}{\partial u_1}\frac{\partial}{\partial u_2}\left\{e^{-G_0(C)\psi_b(u_1,u_2)}\right\}e^{-G_0(C^c)\psi_b(u_1,u_2)}\,\mathrm{d}u_1\mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \frac{\partial}{\partial u_1}\left\{-G_0(C)\frac{\partial}{\partial u_2}\psi_b(u_1,u_2)e^{-G_0(C)\psi_b(u_1,u_2)}\right\}e^{-G_0(C^c)\psi_b(u_1,u_2)}\,\mathrm{d}u_1\mathrm{d}u_2.
\end{aligned}$$

Performing the derivative with respect to u_1, the latter expression can be written as follows

$$
\begin{aligned}
&= \int_{\mathbb{R}_+^2} \left\{ G_0(C)^2 \frac{\partial}{\partial u_1}\psi_b(u_1,u_2)\frac{\partial}{\partial u_2}\psi_b(u_1,u_2) \right\} e^{-G_0(C)\psi_b(u_1,u_2)} e^{-G_0(C^c)\psi_b(u_1,u_2)}\,\mathrm{d}u_1\mathrm{d}u_2 + \\
&+ \int_{\mathbb{R}_+^2} \left\{ -G_0(C)\frac{\partial}{\partial u_1 \partial u_2}\psi_b(u_1,u_2) \right\} e^{-G_0(C)\psi_b(u_1,u_2)} e^{-G_0(C^c)\psi_b(u_1,u_2)}\,\mathrm{d}u_1\mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \left\{ G_0(C)^2 \frac{\partial}{\partial u_1}\psi_b(u_1,u_2)\frac{\partial}{\partial u_2}\psi_b(u_1,u_2) \right\} e^{-\psi_b(u_1,u_2)}\,\mathrm{d}u_1\mathrm{d}u_2 + \\
&+ \int_{\mathbb{R}_+^2} \left\{ -G_0(C)\frac{\partial}{\partial u_1 \partial u_2}\psi_b(u_1,u_2) \right\} e^{\psi_b(u_1,u_2)}\,\mathrm{d}u_1\mathrm{d}u_2
\end{aligned}
$$

By Lemma 1 we then obtain

$$
\int_{\mathbb{R}_+^2} \mathrm{E}\left(e^{-u_1\mu_1(\mathbb{X}\times\mathbb{X})-u_2\mu_2(\mathbb{X}\times\mathbb{X})}\mu_1(C)\mu_2(C)\right)\,\mathrm{d}u_1\mathrm{d}u_2 = G_0(C)^2\,(1-\delta) + G_0(C)\delta,
$$

as desired. □

Lemma 3. *Let C, D be such that $C \cap D = \emptyset$. Then*

$$
\int_{\mathbb{R}_+^2} E\left(e^{-u_1\mu_1(\mathbb{X}\times\mathbb{X})-u_2\mu_2(\mathbb{X}\times\mathbb{X})}\mu_1(C)\mu_2(D)\right)\,\mathrm{d}u_1\mathrm{d}u_2 = G_0(C)G_0(D)\,(1-\delta)
$$

Proof of Lemma S3. Let $Y = (C \cup D)^c$. Since C and D are disjoint, by independence of evaluations on disjoint sets it holds

$$
\begin{aligned}
&\int_{\mathbb{R}_+^2} \mathrm{E}\left(e^{-u_1\mu_1(\mathbb{X}\times\mathbb{X})-u_2\mu_2(\mathbb{X}\times\mathbb{X})}\mu_1(C)\mu_2(D)\right)\,\mathrm{d}u_1\mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \mathrm{E}\left(e^{-u_1\mu_1(C\cup D)-u_2\mu_2(C\cup D)}\mu_1(C)\mu_2(D)\right\}\mathrm{E}\left\{e^{-u_1\mu_1(Y)-u_2\mu_2(Y)}\right)\,\mathrm{d}u_1\mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \mathrm{E}\left(e^{-u_1\mu_1(C)-u_2\mu_2(C)}\mu_1(C)\right)\mathrm{E}\left(e^{-u_1\mu_1(D)-u_2\mu_2(D)}\mu_2(D)\right)\times \\
&\qquad \times \mathrm{E}\left(e^{-u_1\mu_1(Y)-u_2\mu_2(Y)}\right)\,\mathrm{d}u_1\mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \frac{\partial}{\partial u_1}\left\{e^{-G_0(C)\psi_b(u_1,u_2)}\right\}\frac{\partial}{\partial u_2}\left\{e^{-G_0(D)\psi_b(u_1,u_2)}\right\}e^{-G_0(Y)\psi_b(u_1,u_2)}\,\mathrm{d}u_1\mathrm{d}u_2 \\
&= G_0(C)G_0(D)\int_{\mathbb{R}_+^2}\left\{\frac{\partial}{\partial u_1}\psi_b(u_1,u_2)\frac{\partial}{\partial u_2}\psi_b(u_1,u_2)\right\}e^{-\psi_b(u_1,u_2)}\,\mathrm{d}u_1\mathrm{d}u_2
\end{aligned}
$$

The result follows by applying Lemma 1. □

Proof of Proposition 7. We have

$$\begin{aligned}
&\mathbb{P}\left(X \in A, Y \in B\right) = \mathrm{E}\left(\frac{\tilde{\mu}_1(A)}{\tilde{\mu}_1(\mathbb{X})}\frac{\tilde{\mu}_2(B)}{\tilde{\mu}_2(\mathbb{X})}\right) = \mathrm{E}\left(\frac{\mu_1(A\times\mathbb{X})}{\mu_1(\mathbb{X}\times\mathbb{X})}\frac{\mu_2(\mathbb{X}\times B)}{\mu_2(\mathbb{X}\times\mathbb{X})}\right) = \\
&= \int_{\mathbb{R}^2_+} \mathrm{E}\left(e^{-u_1\mu_1(\mathbb{X}\times\mathbb{X})-u_2\mu_2(\mathbb{X}\times\mathbb{X})}\mu_1(A\times\mathbb{X})\mu_2(\mathbb{X}\times B)\right)\,\mathrm{d}u_1\mathrm{d}u_2 = \\
&= \int_{\mathbb{R}^2_+} \mathrm{E}\Big(e^{-u_1\mu_1(\mathbb{X}\times\mathbb{X})-u_2\mu_2(\mathbb{X}\times\mathbb{X})}\big\{\mu_1(A\times B)\mu_2(A\times B)+\mu_1(A\times B)\mu_2(A^c\times B)+ \\
&+\mu_1(A\times B^c)\mu_2(A\times B)+\mu_1(A\times B^c)\mu_2(A^c\times B)\big\}\Big)\,\mathrm{d}u_1\mathrm{d}u_2
\end{aligned}$$

We compute each integral separately applying Lemmas 2 and 3 and obtain

$$\begin{aligned}
\mathbb{P}\left(X \in A, Y \in B\right) &= G_0(A\times\mathbb{X})G_0(\mathbb{X}\times B)\left(1-\delta\right) + G_0(A\times B)\delta \\
&= P_0(A)P_0(B)\left(1-\delta\right) + G_0(A\times B)\delta,
\end{aligned} \tag{7}$$

as desired. Then the probability of a tie in the product space is given exactly by δ, denoted γ in the manuscript. The probability of a tie is given by the particular case $\psi_b(u_1,u_2)=\psi(u_1+u_2)$, since

$$-\int_{\mathbb{R}^2_+}\left\{\frac{\partial^2}{\partial u_1\partial u_2}\psi_b(u_1+u_2)\right\}e^{-\psi_b(u_1+u_2)}\,\mathrm{d}u_1\mathrm{d}u_2 = -\int_0^\infty\int_0^u \mathrm{d}v\left\{\frac{\partial^2}{\partial u^2}\psi_b(u)\right\}e^{-\psi_b(u)}\,\mathrm{d}u,$$

with the change of variables $u=u_1+u_2$ and $v=u_1$. □

Proof of Proposition 8. Since

$$\mathrm{E}\left(\tilde{p}_1(A)\tilde{p}_2(B)\right) = \mathbb{P}\left(X\in A, Y\in B\right),$$

by (7) we have

$$\mathrm{E}\left(\tilde{p}_1(A)\tilde{p}_2(B)\right) = G_0(A\times\mathbb{X})G_0(\mathbb{X}\times B)\left(1-\gamma\right)+G_0(A\times B)\gamma.$$

Finally,

$$\begin{aligned}
\mathrm{cov}\left(\tilde{p}_1(A),\tilde{p}_2(B)\right) &= G_0(A\times\mathbb{X})G_0(\mathbb{X}\times B)\left(1-\gamma\right)+G_0(A\times B)\gamma - G_0(A\times\mathbb{X})G_0(\mathbb{X}\times B) \\
&= \gamma\left\{G_0(A\times B)-G_0(A\times\mathbb{X})G_0(\mathbb{X}\times B)\right\}.
\end{aligned}$$

From this one also obtains

$$\begin{aligned}
\mathrm{var}\left(\tilde{p}_1(A)\right) = \mathrm{cov}\left(\tilde{p}_1(A),\tilde{p}_1(A)\right) &= \beta\left\{P_0(A)-P_0(A)^2\right\} \\
&= \beta P_0(A)\left\{1-P_0(A)\right\},
\end{aligned}$$

and the desired result follows. □

S1.4 Proofs of Section 5

Proof of Theorem 1. We need to compute the conditional Laplace functional of (μ_1, μ_2), i.e.

$$\mathrm{E}\left(e^{-\int_{\mathbb{X}^2} h_1(x)\,\mu_1(\mathrm{d}x)-\int_{\mathbb{X}^2} h_2(x)\,\mu_2(\mathrm{d}x)} \mid (X_i)_{i=1}^n, (Y_j)_{j=1}^m\right),$$

with $h_i : \mathbb{X}^2 \to \mathbb{R}^+$ measurable functions. Define $A_j = A_{j,\epsilon} = \{x \in \mathbb{X} \mid d(x, X_i^*) < \epsilon\}$ and $B_j = B_{j,\epsilon} = \left\{x \in \mathbb{X} \mid d(x, Y_j^*) < \epsilon\right\}$, with $1 \le i \le k$ and $1 \le j \le c$, such that $A_i \cap A_j = \emptyset$ and $B_i \cap B_j = \emptyset$ for any $i \neq j$. Moreover, denote

$$A_{k+1} = \left(\cup_{i=1}^k A_i\right)^c, \quad B_{c+1} = \left(\cup_{i=1}^c B_i\right)^c.$$

Thus our goal becomes to compute

$$\begin{aligned}
&\mathrm{E}\left(e^{-\int_{\mathbb{X}^2} h_1(x)\,\mu_1(\mathrm{d}x)-\int_{\mathbb{X}^2} h_2(x)\,\mu_2(\mathrm{d}x)} \mid (X_i)_{i=1}^n, (Y_j)_{j=1}^m\right)\\
&= \lim_{\epsilon\to 0} \mathrm{E}\left(e^{-\int_{\mathbb{X}^2} h_1(x)\,\mu_1(\mathrm{d}x)-\int_{\mathbb{X}^2} h_2(x)\,\mu_2(\mathrm{d}x)} \mid \underline{X}_n^* \in \times_{j=1}^k A_j, \underline{Y}_m^* \in \times_{j=1}^c B_j\right)\\
&= \lim_{\epsilon\to 0} \frac{\mathrm{E}\left(e^{-\int_{\mathbb{X}^2} h_1(x)\,\mu_1(\mathrm{d}x)-\int_{\mathbb{X}^2} h_2(x)\,\mu_2(\mathrm{d}x)} \prod_{j=1}^k \tilde{p}_1(A_j)^{n_j} \prod_{j=1}^c \tilde{p}_2(B_j)^{m_j}\right)}{\mathrm{E}\left(\prod_{j=1}^k \tilde{p}_1(A_j)^{n_j} \prod_{j=1}^c \tilde{p}_2(B_j)^{m_j}\right)}.
\end{aligned} \tag{8}$$

We start to evaluate

$$\begin{aligned}
&\mathrm{E}\left(\tilde{p}_1(A_1)^{n_1} \dots \tilde{p}_1(A_k)^{n_k} \tilde{p}_2(B_1)^{m_1} \tilde{p}_2(B_c)^{m_c}\right) =\\
&= \mathrm{E}\left(\frac{\tilde{\mu}_1(A_1)^{n_1} \dots \tilde{\mu}_1(A_k)^{n_k} \tilde{\mu}_2(B_1)^{m_1} \tilde{\mu}_2(B_c)^{m_c}}{\tilde{\mu}_1(\mathbb{X})^n \tilde{\mu}_2(\mathbb{X})^m}\right)\\
&= \mathrm{E}\left(\frac{\mu_1(A_1 \times \mathbb{X})^{n_1} \dots \mu_1(A_k \times \mathbb{X})^{n_k} \mu_2(\mathbb{X} \times B_1)^{m_1} \mu_2(\mathbb{X} \times B_c)^{m_c}}{\mu_1(\mathbb{X} \times \mathbb{X})^n \mu_2(\mathbb{X} \times \mathbb{X})^m}\right) = \mathcal{I}.
\end{aligned}$$

By Netwon's binomial

$$\mu_1(A_h \times \mathbb{X}) = \sum_{i_1^h + \dots i_{c+1}^h = n_h} \binom{n_h}{i_1^h, \dots, i_{c+1}^h} \prod_{r=1}^{c+1} \mu_1^{i_r^h}(A_h \times B_r), \quad h = 1, \dots, k,$$

$$\mu_2(\mathbb{X} \times B_r) = \sum_{j_1^r + \dots j_{k+1}^r = m_r} \binom{m_r}{j_1^r, \dots, j_{k+1}^r} \prod_{h=1}^{k+1} \mu_2^{j_h^r}(A_h \times B_r), \quad r = 1, \dots, c.$$

For ease of notation denote

$$\sum_{\boldsymbol{i},\boldsymbol{j}} \binom{\boldsymbol{n}}{\boldsymbol{i}} \binom{\boldsymbol{m}}{\boldsymbol{j}} = \sum_{i_1^1+\ldots i_{c+1}^1=n_1} \binom{n_1}{i_1^1,\ldots,i_{c+1}^1} \cdots \sum_{i_1^{c+1}+\ldots i_{c+1}^{k+1}=n_{k+1}} \binom{n_{k+1}}{i_1^{k+1},\ldots,i_{c+1}^{k+1}} \times$$
$$\times \sum_{j_1^1+\ldots j_{k+1}^1=m_1} \binom{m_1}{j_1^1,\ldots,j_{k+1}^1} \cdots \sum_{j_1^{k+1}+\ldots j_{k+1}^{k+1}=m_{k+1}} \binom{m_{k+1}}{j_1^{k+1},\ldots,j_{k+1}^{k+1}}.$$

Thus

$$\mathcal{I} = \sum_{\boldsymbol{i},\boldsymbol{j}} \binom{\boldsymbol{n}}{\boldsymbol{i}} \binom{\boldsymbol{m}}{\boldsymbol{j}} \mathcal{I}_{\boldsymbol{i},\boldsymbol{j}},$$

with

$$\mathcal{I}_{\boldsymbol{i},\boldsymbol{j}} = \mathrm{E}\Bigg(\frac{\prod_{h=1}^k \prod_{r=1}^c \mu_1^{i_r^h}(A_h \times B_r) \mu_2^{j_h^r}(A_h \times B_r)}{\mu_1(\mathbb{X}\times\mathbb{X})^n} \times$$
$$\times \frac{\prod_{h=1}^k \mu_1^{i_{c+1}^h}(A_h \times B_{c+1}) \prod_{r=1}^c \mu_2^{j_{k+1}^r}(A_{k+1} \times B_r)}{\mu_2(\mathbb{X}\times\mathbb{X})^m} \Bigg)$$

Letting $\mu_1 := \mu_1(\mathbb{X}\times\mathbb{X})$ and $\mu_2 := \mu_2(\mathbb{X}\times\mathbb{X})$, we have

$$\frac{1}{\mu_1(\mathbb{X}\times\mathbb{X})^n \mu_2(\mathbb{X}\times\mathbb{X})^m} = \frac{1}{\Gamma(n)\Gamma(m)} \int_{\mathbb{R}_+^2} u_1^{n-1} u_2^{m-1} e^{-u_1\mu_1 - u_2\mu_2} \, \mathrm{d}\underline{u},$$

with $\underline{u} = (u_1, u_2)$. Thus, by Fubini's Theorem

$$\mathcal{I}_{\mathbf{i},\mathbf{j}} = \int_{\mathbb{R}_+^2} \frac{u_1^{n-1} u_2^{m-1}}{\Gamma(n)\Gamma(m)} \mathrm{E}\Bigg(e^{-u_1\mu_1 - u_2\mu_2} \left\{ \prod_{h=1}^k \prod_{r=1}^c \mu_1^{i_r^h}(A_h \times B_r) \mu_2^{j_h^r}(A_h \times B_r) \right\} \times$$
$$\times \prod_{h=1}^k \mu_1^{i_{c+1}^h}(A_h \times B_{c+1}) \prod_{r=1}^c \mu_2^{j_{k+1}^r}(A_{k+1} \times B_r) \Bigg) \, \mathrm{d}\underline{u} =$$
$$= \int_{\mathbb{R}_+^2} \frac{u_1^{n-1} u_2^{m-1}}{\Gamma(n)\Gamma(m)} \rho_{\mathbf{i},\mathbf{j}}(\underline{u}) \, \mathrm{d}\underline{u}.$$

By independence of evaluations on disjoint sets we have

$$\rho_{\mathbf{i},\mathbf{j}}(\underline{u}) = \mathrm{E}\Bigg(\left\{ \prod_{h=1}^k \prod_{r=1}^c e^{-u_1\mu_1(A_h\times B_r) - u_2\mu_2(A_h\times B_r)} \mu_1^{i_r^h}(A_h \times B_r) \mu_2^{j_h^r}(A_h \times B_r) \right\} \times$$
$$\times \left\{ \prod_{h=1}^k e^{-u_1\mu_1(A_h\times B_{c+1}) - u_2\mu_2(A_h\times B_{c+1})} \mu_1^{i_{c+1}^h}(A_h \times B_{c+1}) \right\} \times$$
$$\times \left\{ \prod_{r=1}^c e^{-u_1\mu_1(A_{k+1}\times B_r) - u_2\mu_2(A_{k+1}\times B_r)} \mu_2^{j_{k+1}^r}(A_{k+1} \times B_r) \right\} \Bigg)$$

This can be equivalently written as

$$\prod_{h=1}^{k}\prod_{r=1}^{c}\mathrm{E}\left(e^{-u_1\mu_1(A_h\times B_r)-u_2\mu_2(A_h\times B_r)}\mu_1^{i_r^h}(A_h\times B_r)\mu_2^{j_h^r}(A_h\times B_r)\right)\times$$
$$\times\prod_{h=1}^{k}\mathrm{E}\left(e^{-u_1\mu_1(A_h\times B_{c+1})-u_2\mu_2(A_h\times B_{c+1})}\mu_1^{i_{c+1}^h}(A_h\times B_{c+1})\right)\times$$
$$\times\prod_{r=1}^{c}\mathrm{E}\left(e^{-u_1\mu_1(A_{k+1}\times B_r)-u_2\mu_2(A_{k+1}\times B_r)}\mu_2^{j_{k+1}^r}(A_{k+1}\times B_r)\right).$$

Considering each element separately we have

$$\begin{aligned}\mathrm{E}\Big(e^{-u_1\mu_1(A_h\times B_r)-u_2\mu_2(A_h\times B_r)}\mu_1^i(A_h\times B_r)\mu_2^j(A_h\times B_r)\Big)&\\ =\mathrm{E}\left((-1)^{i+j}\frac{\partial^{i+j}}{\partial u_1^i\partial u_2^j}e^{-u_1\mu_1(A_h\times B_r)-u_2\mu_2(A_h\times B_r)}\right)&\\ =(-1)^{i+j}\frac{\partial^{i+j}}{\partial u_1^i\partial u_2^j}\mathrm{E}\left(e^{-u_1\mu_1(A_h\times B_r)-u_2\mu_2(A_h\times B_r)}\right)&\\ =(-1)^{i+j}\frac{\partial^{i+j}}{\partial u_1^i\partial u_2^j}\left\{e^{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(ds)G_0(x)}\right\}.&\end{aligned}$$

Recall that we are interested in the limit as $\epsilon\to 0$, so that

$$\begin{aligned}\frac{\partial^{i+j}}{\partial u_1^i\partial u_2^j}\left\{e^{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}\right\}&\sim e^{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}\times\\ \times\frac{\partial^{i+j}}{\partial u_1^i\partial u_2^j}\left\{\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right\},&\end{aligned}\tag{9}$$

where we say $f\sim g$ if $\lim_{\epsilon\to 0}f(x)/g(x)=1$. By simple algebra we get

$$\begin{aligned}\frac{\partial^{i+j}}{\partial u_1^i\partial u_2^j}\left\{e^{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}\right\}&=\frac{\partial^{i+j-1}}{\partial u_1^{i-1}\partial u_2^j}\Bigg\{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}e^{-u_1s_1-u_2s_2}\times\\ &\times s_1\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)e^{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}\Bigg\}\\ &=\frac{\partial^{i+j-2}}{\partial u_1^{i-2}\partial u_2^j}\Bigg\{\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}e^{-u_1s_1-u_2s_2}s_1^2\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)e^{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}\\ &+\left(\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}e^{-u_1s_1-u_2s_2}s_1\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right)^2e^{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}\Bigg\},\end{aligned}$$

and

$$\lim_{\epsilon\to 0}\frac{\left(\int_{A_h\times B_r}\int_{\mathbb{R}_+^2} e^{-u_1s_1-u_2s_2}s_1\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right)^2}{\int_{A_h\times B_r}\int_{\mathbb{R}_+^2} e^{-u_1s_1-u_2s_2}s_1^2\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}=0.$$

By applying this argument repeatedly we obtain (9). Thus, letting $\rho(\underline{u})=\sum_{\boldsymbol{i},\boldsymbol{j}}\binom{\boldsymbol{n}}{\boldsymbol{i}}\binom{\boldsymbol{m}}{\boldsymbol{j}}\rho_{\boldsymbol{i},\boldsymbol{j}}(\underline{u})$, by aggregating the terms we have

$$\begin{aligned}
\rho(\underline{u}) \sim &\sum_{\boldsymbol{i},\boldsymbol{j}}\binom{\boldsymbol{n}}{\boldsymbol{i}}\binom{\boldsymbol{m}}{\boldsymbol{j}}(-1)^{n+m}e^{-\psi_b(u)}\times\\
&\times\prod_{h=1}^{k}\prod_{r=1}^{c}\left\{\frac{\partial^{i_r^h+j_h^r}}{\partial u_1^{i_r^h}\partial u_2^{j_h^r}}\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right\}\times\\
&\times\prod_{h=1}^{k}\left\{\frac{\partial^{i_{c+1}^h}}{\partial u_1^{i_{c+1}^h}}\int_{A_h\times B_{c+1}}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right\}\times\\
&\times\prod_{r=1}^{c}\left\{\frac{\partial^{j_{k+1}^r}}{\partial u_2^{i_{k+1}^r}}\int_{A_{k+1}\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right\}\\
&=\sum_{\boldsymbol{i},\boldsymbol{j}}\binom{\boldsymbol{n}}{\boldsymbol{i}}\binom{\boldsymbol{m}}{\boldsymbol{j}}(-1)^{n+m}V(\boldsymbol{i},\boldsymbol{j}).
\end{aligned}$$

The following three Lemmas characterize the set of indices $(\boldsymbol{i},\boldsymbol{j})$ that are relevant once the limit is taken.

Lemma 4. *Consider $(\boldsymbol{i},\boldsymbol{j})$ such that $0<i_r^h,i_l^h<n_h$, with $r>l$ and $1\le h\le k$. Then $\exists(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}})$ such that $\lim_{\epsilon\to 0}V(\boldsymbol{i},\boldsymbol{j})/V(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}})\to 0$.*

Proof of Lemma S2.4. For ease of notation set $\boldsymbol{i}^h=(i_1^h,\dots,i_{c+1}^h)$. Then

- If $r=c+1$, set $\tilde{\boldsymbol{i}}^h=(i_1^h,\dots,i_l^h+i_{c+1}^h,\dots,0)$.
- If $j_h^r=0$, set $\tilde{i}^h=(i_1^h,\dots,i_l^h+i_r^h,\dots,0,\dots)$.
- If $j_h^l=0$, set $\tilde{\boldsymbol{i}}^h=(i_1^h,\dots,0,\dots,i_r^h+i_l^h,\dots)$.
- If $j_h^l>0$ and $j_h^r>0$, set $\tilde{\boldsymbol{j}}^r=(j_1^r,\dots,0,\dots,j_{k+1}^r+j_h^r)$ and $\tilde{\boldsymbol{i}}^h=(i_1^h,\dots,i_l^h+i_r^h,\dots,0,\dots)$.

For example in the last case we have

$$\lim_{\epsilon\to 0}\frac{\mathrm{var}(\boldsymbol{i},\boldsymbol{j})}{\mathrm{var}(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}})}=\lim_{\epsilon\to 0}\frac{\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}e^{-u_1s_1-u_2s_2}s_1^{i_r^h}s_2^{j_h^r}\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}{\int_{A_{c+1}\times B_r}\int_{\mathbb{R}_+^2}e^{-u_1s_1-u_2s_2}s_2^{j_h^r+j_{c+1}^r}\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}=0,$$

as desired. □

Thus, Lemma 4 guarantees that $\boldsymbol{i}^h$ has exactly one element different from 0, that is equal to n_h.

Lemma 5. *Consider* $(\boldsymbol{i},\boldsymbol{j})$ *such that* $i_r^h = n_h$ *and* $j_h^r = 0$*. Then there exists* $(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}})$ *such that*

$$\lim_{\epsilon\to 0} V(\boldsymbol{i},\boldsymbol{j})/V(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}}) \to 0.$$

Proof of Lemma S5. Set $(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}})$ equal to $(\mathbf{i},\mathbf{j})$, apart from $\tilde{i}_r^h = 0$ and $\tilde{i}_{c+1}^h = n_h$. □

Lemma 6. *Consider* $(\boldsymbol{i},\boldsymbol{j})$ *such that* $i_{c+1}^h = n_h$ *and* $j_h^r > 0$*. Then there exists* $(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}})$ *such that*

$$\lim_{\epsilon\to 0} V(\boldsymbol{i},\boldsymbol{j})/V(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}}) \to 0.$$

Proof of Lemma S2.6. Set $(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}})$ equal to $(\boldsymbol{i},\boldsymbol{j})$, apart from $\tilde{j}_h^r = 0$ and $\tilde{j}_{k+1}^r = m_r$. □

The three lemmas imply that each relevant $(\boldsymbol{i},\boldsymbol{j})$ corresponds to an admissible latent structure, i.e.

$$\begin{aligned}\rho(u) \sim \sum_{\mathbf{p}\in\mathcal{P}} (-1)^{n+m} e^{-\psi_b(\underline{u})} &\prod_{(i,j)\in\Delta_{\mathbf{p}}} \left\{ \frac{\partial^{n_i+m_j}}{\partial u_1^{n_i}\partial u_2^{m_j}} \int_{A_i\times B_j}\int_{\mathbb{R}_+^2} (1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right\}\times\\ &\times \prod_{(i,j)\in\Delta_{\mathbf{p}}^1} \left\{ \frac{\partial^{n_i}}{\partial u_1^{n_i}} \int_{A_i\times B_{c+1}}\int_{\mathbb{R}_+^2} (1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right\}\times\\ &\times \prod_{(i,j)\in\Delta_{\mathbf{p}}^2} \left\{ \frac{\partial^{m_j}}{\partial u_2^{m_j}} \int_{A_{k+1}\times B_j}\int_{\mathbb{R}_+^2} (1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right\}.\end{aligned}$$

Evaluating the derivatives we have

$$\begin{aligned}\rho(\underline{u}) \sim \sum_{\mathbf{p}\in\mathcal{P}} e^{-\psi_b(\underline{u})} &\prod_{(i,j)\in\Delta_{\mathbf{p}}} \left\{ \int_{A_i\times B_j}\int_{\mathbb{R}_+^2} e^{-u_1s_1-u_2s_2} s_1^{n_i} s_2^{m_j}\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right\}\times\\ &\times \prod_{(i,j)\in\Delta_{\mathbf{p}}^1} \left\{ \int_{A_i\times B_{c+1}}\int_{\mathbb{R}_+^2} e^{-u_1s_1-u_2s_2} s_1^{n_i}\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right\}\times\\ &\times \prod_{(i,j)\in\Delta_{\mathbf{p}}^2} \left\{ \int_{A_{k+1}\times B_j}\int_{\mathbb{R}_+^2} e^{-u_1s_1-u_2s_2} s_2^{m_j}\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right\}.\end{aligned}$$

Finally, we get

$$\mathcal{I} \sim \sum_{\mathbf{p}\in\mathcal{P}} \int_{\mathbb{R}_+^2} \frac{u_1^{n-1}u_2^{m-1}}{\Gamma(n)\Gamma(m)} e^{-\psi_b(\underline{u})} \prod_{(i,j)\in\Delta_{\mathbf{p}}} \left\{ \int_{A_i\times B_j} \int_{\mathbb{R}_+^2} e^{-u_1 s_1 - u_2 s_2} s_1^{n_i} s_2^{m_j} \, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \times$$
$$\times \prod_{(i,j)\in\Delta_{\mathbf{p}}^1} \left\{ \int_{A_i\times B_{c+1}} \int_{\mathbb{R}_+^2} e^{-u_1 s_1 - u_2 s_2} s_1^{n_i} \, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \times$$
$$\times \prod_{(i,j)\in\Delta_{\mathbf{p}}^2} \left\{ \int_{A_{k+1}\times B_j} \int_{\mathbb{R}_+^2} e^{-u_1 s_1 - u_2 s_2} s_2^{m_j} \, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \mathrm{d}u.$$

Evaluating the numerator of (8) the same reasoning yields a formula asymptotic to

$$\sum_{\mathbf{p}\in\mathcal{P}} \int_{\mathbb{R}_+^2} \frac{u_1^{n-1}u_2^{m-1}}{\Gamma(n)\Gamma(m)} e^{-\psi_h(\underline{u})} \prod_{(i,j)\in\Delta_{\mathbf{p}}} \left\{ \int_{A_i\times B_j} \int_{\mathbb{R}_+^2} e^{-(h_1(x)+u_1)s_1 - (h_2(x)+u_2)s_2} s_1^{n_i} s_2^{m_j} \, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\}$$
$$\prod_{(i,j)\in\Delta_{\mathbf{p}}^1} \left\{ \int_{A_i\times B_{c+1}} \int_{\mathbb{R}_+^2} e^{-(h_1(x)+u_1)s_1 - (h_2(x)+u_2)s_2} s_1^{n_i} \, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\}$$
$$\prod_{(i,j)\in\Delta_{\mathbf{p}}^2} \left\{ \int_{A_{k+1}\times B_j} \int_{\mathbb{R}_+^2} e^{-(h_1(x)+u_1)s_1 - (h_2(x)+u_2)s_2} s_2^{m_j} \, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \mathrm{d}u.$$

where $\psi_h(\underline{u}) = \int_{\mathbb{X}^2} \int_{\mathbb{R}_+^2} \left(1 - e^{-(h_1(x)+u_1)s_1 - (h_2(x)+u_2)s_2}\right) \rho(\mathrm{d}s) G_0(\mathrm{d}x)$. Note that

$$\begin{aligned} 1 - e^{-(h_1(x)+u_1)s_1 - (h_2(x)+u_2)s_2} &= e^{-u_1 s_1 - u_2 s_2} \left[e^{u_1 s_1 + u_2 s_2} - 1 + 1 - e^{-h_1(x)s_1 - h_2(x)s_2} \right] \\ &= \left[1 - e^{-u_1 s_1 - u_2 s_2}\right] + \left[1 - e^{-h_1(x)s_1 - h_2(x)s_2}\right], \end{aligned}$$

so that

$$\begin{aligned} e^{-\psi_h(\underline{u})} &= e^{-\psi_b(\underline{u})} e^{-\int_{\mathbb{X}^2}\int_{\mathbb{R}_+^2} \left[1 - e^{-h_1(x)s_1 - h_2(x)s_2}\right] \rho(\mathrm{d}s) G_0(\mathrm{d}x)} \\ &= e^{-\psi_b(\underline{u})} \mathrm{E}\left[e^{-\int_{\mathbb{X}^2} h_1(x)\, \hat{\mu}_1(\mathrm{d}x) - \int_{\mathbb{X}^2} h_2(x)\, \hat{\mu}_2(\mathrm{d}x)} \right]. \end{aligned}$$

Furthermore

$$G_0(A_h \times B_r) = \epsilon \frac{G_0(A_h \times B_r)}{\epsilon} \sim \epsilon g_{h,r}, \quad 1 \le i \le c, 1 \le j \le k,$$

and

$$G_0(A_h \times \mathrm{d}x) \sim \epsilon g_{h,c+1} Q_{X_h^*}(\mathrm{d}x), \quad G_0(\mathrm{d}x \times B_r) \sim \epsilon g_{k+1,r} P_{Y_r^*}(\mathrm{d}x).$$

Thus, evaluating the limit in (8) we get

$$
\begin{aligned}
&\mathrm{E}\left[e^{-\int_{\mathbb{X}^2} h_1(x)\,\mu_1(\mathrm{d}x)-\int_{\mathbb{X}^2} h_2(x)\,\mu_2(\mathrm{d}x)} \mid (X_i)_{i\geq 1}^n, (Y_j)_{j\geq 1}^m\right] = \\
&\quad \times \sum_{p\in\mathcal{P}} \int_{\mathbb{R}_+^2} \mathrm{E}\left[e^{-\int_{\mathbb{X}^2} h_1(x)\,\hat{\mu}_1(\mathrm{d}x)-\int_{\mathbb{X}^2} h_2(x)\,\hat{\mu}_2(\mathrm{d}x)}\right] \times \\
&\quad \times \prod_{(i,j)\in\Delta_p} \int_{\mathbb{R}_+^2} e^{-h_1(X_i^*,Y_j^*)s_1-h_2(X_i^*,Y_j^*)s_2} \frac{s_1^{n_i} s_2^{m_j} e^{-u_1 s_1-u_2 s_2}\rho(\mathrm{d}s)}{\tau_{n_i,m_j}(\underline{u})} \times \\
&\quad \times \prod_{(i,j)\in\Delta_p^1} \int_{\mathbb{X}}\int_{\mathbb{R}_+^2} e^{-h_1(X_i^*,x_2)s_1-h_2(X_i^*,x_2)s_2} \frac{s_1^{n_i} e^{-u_1 s_1-u_2 s_2}\rho(\mathrm{d}s)}{\tau_{n_i,0}(\underline{u})} Q_{X_i^*}(\mathrm{d}x_2) \times \\
&\quad \times \prod_{(i,j)\in\Delta_p^2} \int_{\mathbb{X}}\int_{\mathbb{R}_+^2} e^{-h_1(x_1,Y_2^*)s_1-h_2(x_1,Y_2^*)s_2} \frac{s_2^{m_j} e^{-u_1 s_1-u_2 s_2}\rho(\mathrm{d}s)}{\tau_{0,m_j}(\underline{u})} P_{Y_j^*}(\mathrm{d}x_1) \times \\
&\quad \times \left(\frac{\int_{\mathbb{R}_+^2} u_1^{n-1}u_2^{m-1}\prod_{(i,j)\in p} g_{i,j}\tau_{n_i,m_j}(\underline{u})e^{-\psi_b(\underline{u})}\,\mathrm{d}\underline{u}}{\sum_{\mathbf{q}\in\mathcal{P}}\int_{\mathbb{R}_+^2} u_1^{n-1}u_2^{m-1}\prod_{(i,j)\in\mathbf{q}} g_{i,j}\tau_{n_i,m_j}(\underline{u})e^{-\psi_b(\underline{u})}\mathrm{d}\underline{u}}\right) \times \\
&\quad \times \frac{u_1^{n-1}u_2^{m-1}\prod_{(i,j)\in p}\tau_{n_i,m_j}(\underline{u})e^{-\psi_b(\underline{u})}\,\mathrm{d}u}{\int_{\mathbb{R}_+^2} u_1^{n-1}u_2^{m-1}\prod_{(i,j)\in p}\tau_{n_i,m_j}(\underline{u})e^{-\psi_b(\underline{u})}\,\mathrm{d}\underline{u}},
\end{aligned}
$$

as desired. □

Proof of Corollary 2. We use the shorthand notation $\mu_1(f) = \int_{\mathbb{X}} f(x)\,\mu_1(\mathrm{d}x)$ for any measurable function $f : \mathbb{X} \to \mathbb{R}$ such that $\mu_1(|f|) < \infty$. Letting U be the set of latent variables of Theorem 1, i.e. $U = (p, U_1, U_2, Z^x, Z^y)$ for any $y_1, \ldots, y_n \in (0,1)$ and $A_1, \ldots, A_n \in \mathcal{X}^2$ we get

$$
\begin{aligned}
&\mathbb{P}\left[p_1(A_1) \leq y_1, \ldots, p_n(A_n) \leq y_n \mid U, (X_i)_{i=1}^n, (Y_j)_{j=1}^m\right] \\
=&\mathbb{P}\left[\mu_1(\mathbb{1}_{A_1} - y_1) \leq 0, \ldots, \mu_n(\mathbb{1}_{A_n} - y_n) \leq 0 \mid U, (X_i)_{i=1}^n, (Y_j)_{j=1}^m\right].
\end{aligned}
$$

The result follows since the finite dimensional distributions of p_1 given U, $(X_i)_{i=1}^n$, and $(Y_j)_{j=1}^m$ coincide with the ones of the normalized posterior distribution of μ_1, given U, $(X_i)_{i=1}^n$, and $(Y_j)_{j=1}^m$. □

Proof of Theorem 2. Set $\underline{H} = (p, U_1, U_2)$ with domain D. Then

$$
\begin{aligned}
\mathbb{P}(X_{n+1} \in \mathrm{d}x \mid (X_i)_{i=1}^n, (Y_j)_{j=1}^m) &= \mathrm{E}[\tilde{p}_1(\mathrm{d}x) \mid (X_i)_{i=1}^n, (Y_j)_{j=1}^m] \\
&= \int_D \mathrm{E}[\tilde{p}_1(\mathrm{d}x) \mid \underline{H} = \underline{h}, (X_i)_{i=1}^n, (Y_j)_{j=1}^m]\, F(\mathrm{d}\underline{v}),
\end{aligned}
$$

where $F(\cdot)$ is the posterior distribution of $\underline{H}$, with $\underline{h} = (p, u_1, u_2)$. Recalling the notation in Corollary 2 we have

$$\mathrm{E}[\tilde{p}_1(\mathrm{d}x) \mid \underline{H} = \underline{h}, (X_i)_{i=1}^n, (Y_j)_{j=1}^m] = \mathrm{E}\left[\frac{\hat{\mu}_1(\mathrm{d}x \times \mathbb{X})}{R}\right] + \mathrm{E}\left[\frac{\sum_{(i,j)\in\Delta_p} J_{i,j}^1 \delta_{X_i^*}}{R}\right] + \\ + \mathrm{E}\left[\frac{\sum_{(i,j)\in\Delta_p^1} J_{i,c+1}^1 \delta_{X_i^*}}{R}\right] + \mathrm{E}\left[\frac{\sum_{(i,j)\in\Delta_p^2} J_{k+1,j}^1 \delta_{Z_j^y}}{R}\right] = \sum_{k=1}^4 I_k,$$

where $R = T_1 + \sum_{(i,j)\in\Delta_p} J_{i,j}^1 + \sum_{(i,j)\in\Delta_p^1} J_{i,c+1}^1 + \sum_{(i,j)\in\Delta_p^2} J_{k+1,j}^1$.
Set $S = \sum_{(i,j)\in\Delta_p} J_{i,j}^1 + \sum_{(i,j)\in\Delta_p^1} J_{i,c+1}^1 + \sum_{(i,j)\in\Delta_p^2} J_{k+1,j}^1$ and exploit the conditional independence between J_{ij}^1 and $\hat{\mu}_1$ to obtain

$$\begin{aligned} I_1 &= \int_{\mathbb{R}_+} \mathrm{E}\left[e^{-vS}\right] \mathrm{E}\left[\hat{\mu}_1(\mathrm{d}x \times \mathbb{X}) e^{-vT_1}\right] \mathrm{d}v \\ &= \theta P_0(\mathrm{d}x) \int_{\mathbb{R}_+} \left(\prod_{(i,j)\in p} \frac{\tau_{n_i,m_j}(u_1+v, u_2)}{\tau_{n_i,m_j}(u_1,u_2)} \right) \tau_{1,0}(u_1+v,u_2) e^{-\psi_b^{\underline{u}}(v,0)} \mathrm{d}v, \end{aligned}$$

where $\psi_b^{\underline{u}}(\lambda_1, \lambda_2)$ is the Laplace exponent of $(\hat{\mu}_1, \hat{\mu}_2)$ in Theorem 1. Observing that $\psi_b^{\underline{u}}(v,0) + \psi(u_1,u_2) = \psi(u_1+v,u_2)$ and denoting with $L(\cdot)$ the distribution of $\mathbf{p}$, we obtain

$$\begin{aligned} \xi_0 &= \int_D I_1 \, F(\mathrm{d}\underline{u}) \\ &= \theta P_0(\mathrm{d}x) \int\int\int_{\mathbb{R}_+^3} \left\{ u_1^{n-1} u_2^{m-1} \left(\prod_{(i,j)\in p} \tau_{n_i,m_j}(u_1+v,u_2) \right) \tau_{1,0}(u_1+v,u_2) \times \right. \\ &\left. \times e^{-\psi(u_1+v,u_2)} \mathrm{d}u_1 \mathrm{d}u_2 \mathrm{d}v L(\mathrm{d}p) \right\} \\ &= \frac{\theta P_0(\mathrm{d}x)}{n} \int\int_{\mathbb{R}_+^2} u_1^n u_2^{m-1} \left(\prod_{(i,j)\in p} \tau_{n_i,m_j}(u_1,u_2) \right) \tau_{1,0}(u_1,u_2) e^{-\psi(u_1,u_2)} \mathrm{d}u_1 \mathrm{d}u_2 L(\mathrm{d}p) \\ &= \frac{\theta P_0(\mathrm{d}x)}{n} \int_D u_1 \tau_{1,0}(u_1,u_2) \, F(\mathrm{d}\underline{u}), \end{aligned}$$

where the second equality follows from the change of variables $(w,z) = (u_1+v, u_1)$. The proof for the remaining weights follows along the same lines and leads to

$$\xi_i^x = \frac{1}{n} \int_D u_1 \left[\frac{\tau_{n_i+1,m_j}(u_1,u_2)}{\tau_{n_i,m_j}(u_1,u_2)} + \frac{\tau_{n_i+1,0}(u_1,u_2)}{\tau_{n_i,0}(u_1,u_2)} \right] F(\mathrm{d}\underline{u})$$

and

$$\xi_i^y = \frac{1}{n} \int_D u_1 \frac{\tau_{1,m_j}(u_1,u_2)}{\tau_{0,m_j}(u_1,u_2)} \, F(\mathrm{d}\underline{u}).$$

The weights for Y_{m+1} can be computed in an analogous fashion. □

S2 A toy example of borrowing of information

Classical borrowing of information across samples is typically associated to positive correlation across observations in different populations and, as a consequence, it induces shrinkage of the predictions. Let us consider the toy situation in which observations coming from two different populations have been collected and a normal model is assumed

$$\begin{aligned} X_i \mid \mu_x &\overset{iid}{\sim} \mathrm{N}(\mu_x, 1) \qquad \text{for } i = 1, \dots, n \\ Y_j \mid \mu_y &\overset{iid}{\sim} \mathrm{N}(\mu_y, 1) \qquad \text{for } j = 1, \dots, m \end{aligned}$$

To obtain a working model, one has to specify a certain prior over μ_x and μ_y. The main typical strategies one may employ are the following:

- Modeling μ_x and μ_y as independent, which ultimately means that we do not consider the information coming from one population to be relevant for inference on the other.
- Modeling μ_x and μ_y as dependent, which induces borrowing of information. This typically reflects the idea that, if the observed values of $Y_1, \dots, Y_m$ are on average higher than our prior guess on μ_y, then we should upwards revise our belief on μ_x and our prediction for X_1.

To clarify this last point, we compare a typical strategy used to perform borrowing of information, which is provided by the following hierarchy

$$\begin{aligned} \mu_x \mid \mu_0 &\sim \mathrm{N}(\mu_0, 1) \\ \mu_y \mid \mu_0 &\sim \mathrm{N}(\mu_0, 1) \\ \mu_0 &\sim \mathrm{N}(\nu, 1) \end{aligned} \tag{10}$$

with the case of independent priors, namely

$$\begin{gathered} \mu_x \sim \mathrm{N}(\nu, 2) \qquad \mu_y \sim \mathrm{N}(\nu, 2) \\ \mu_x \perp \mu_y \end{gathered} \tag{11}$$

where the variance is chosen to match the marginal distributions of the hierarchical specification. We assume that only the sample $(Y_1, \dots, Y_m)$ has been observed and we discuss its impact on the posterior distribution of μ_x and on the predictive distribution of X_1 under the two specifications. Under indepen-

dence in (11), one obviously has

$$p(\mu_x \mid (Y_j)_{j=1}^m) = \mathrm{N}(\nu,\, 2)$$

while under model (10) the new distribution of μ_x is

$$\begin{aligned} p(\mu_x \mid (Y_j)_{j=1}^m) &\propto \int_{\mathbb{R}} p(\mu_x \mid \mu_0)\, p(\mu_0 \mid (Y_j)_{j=1}^m) \mathrm{d}\mu_0 \\ &= \mathrm{N}\left(\frac{1}{2m+1}\nu + \frac{2m}{2m+1}\frac{\nu+\bar{y}}{2},\, 1 + \frac{m+1}{2m+1} \right), \end{aligned}$$

where $\bar{y}$ denotes the empirical average of $Y_1, \ldots, Y_m$, and

$$\mathbb{E}[X_1 \mid (Y_j)_{j=1}^m] = \mathbb{E}[\mu_x \mid (Y_j)_{j=1}^m] = \nu + \frac{m}{2m+1}(\bar{y} - \nu)$$

Therefore, when $\bar{y} > \nu$ the borrowing results in an increase of the estimate for μ_x and of the prediction for X_1, while if $\bar{y} < \nu$ the borrowing of information induces the opposite effect. The shrinking behaviour is ultimately a consequence of the fact that the hierarchical prior in (10) induces positive correlation across X_i and Y_j. However, what we show in the main paper is that classical shrinkage of the estimates is not the only way to borrow information within partially exchangeable populations, neither necessarily the best one.

S3 Example of correlation between FuRBI priors on Borel set

Consider a pair of n-FuRBI priors with equal jumps (see Example 4 in the main document), where the baseline distribution G_0 is given by a bivariate normal with zero mean, unit variances and correlation $\rho \in \{-0.99, -0.5, 0, 0.5, 0.99\}$. In Figure S1 we depict the correlations on sets of the form $(-\infty, x]$, with $x \in [-5, 5]$ and for each value of the correlation. Notice that such correlation may be of particular interest in survival settings, where the distribution function is often the main focus.

When $\rho = 0$, the correlation is equal to 0 as expected, since $G_0(A \times A) = P_0(A)^2$ and the numerator of the formula in Proposition 8 vanishes. For values of ρ different from 0, the correlation is symmetric around 0, due to the symmetry of the Gaussian distribution, and different signs indicate opposite behaviours: therefore, $\rho < 0$ implies negative correlation on such Borel sets.

However, note that a different sign does not mean a completely specular behaviour: for instance the correlation with $\rho = 0.99$ is higher in absolute value than the one with $\rho = -0.99$. This is due to the fact that it is somewhat impossible to have strictly negative correlation on all Borel sets. Intuitively, if the two priors have high negative correlation on $(-\infty, 0]$, it means that one of them has much larger mass on $(-\infty, 0]$ and the other on $(0, +\infty)$: therefore, both priors will have a high mass on $(-\infty, a]$, with a large positive number, so that the correlation can not attain again large negative values.

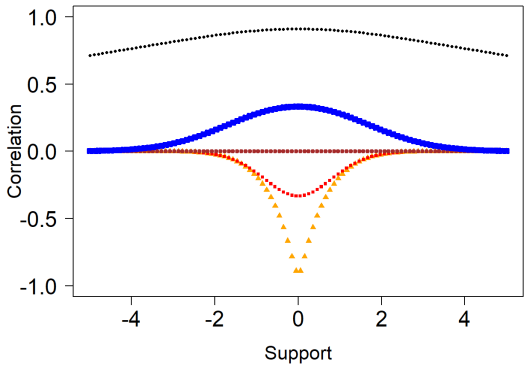


Figure S1: Correlation on Borel sets of the form $(-\infty, x]$, with $x \in [-5, 5]$. The four lines, from bottom to top, correspond to $\rho \in \{-0.99, -0.5, 0, 0.5, 0.99\}$.

Finally, if $\rho \to 1$, then the correlation converges to the constant function 1, that is the value obtained with equal atoms: indeed, the two priors will have equal jumps and linearly dependent atoms (see Corollary 1).

S4 Algorithms for posterior inference

In this section we address the issue of sampling from the posterior distribution. In discrete nonparametric models, we need to distinguish whether the random probability measures are directly applied to the data or rather convoluted with a suitable kernel (known as *mixture* model, see Section 6 in the paper).

Nevertheless, from a computational perspective, if the first problem is solved the second one can be tackled in a similar way: it is indeed easy to propose a Gibbs sampler that alternates sampling of suitable latent variables and the posterior distribution given data originated by the random probability measure (see Section S4.4 below for how to extend algorithms to mixture models).

Therefore, in the following three sections, we assume to collect observations from

$$(X_i, Y_j) \mid (\tilde{p}_1, \tilde{p}_2) \overset{iid}{\sim} \tilde{p}_1 \times \tilde{p}_2 \qquad (\tilde{p}_1, \tilde{p}_2) \sim Q \tag{12}$$

S4.1 Marginal posterior samplers

The first approach is to directly simulate the trajectories of $(\tilde{p}_1, \tilde{p}_2)$ from its posterior, giving rise to so–called conditional algorithms. See, e.g, Ishwaran and James (2001); Walker (2007); Papaspiliopoulos and Roberts (2008); Arbel and Prünster (2017). Conditional samplers for the n-FuRBI priors can be found in Sections S3.2-3 below.

Alternatively, and this is the route followed in this section, one can use marginal algorithms, that integrate

out the random measures and sample sequentially from the predictive distributions (see, for instance, Neal, 2000).

Given $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$ and using the results in Theorem 2, we can sample iteratively new observations from $\tilde{p}_1$ as follows

Marginal algorithm - 1

(a) Compute weights ξ_0, $\{\xi_i^x\}$ and $\{\xi_j^y\}$ from $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$

(b) Draw X_{n+1} from $m(\mathrm{d}x) = \xi_0 P_0(\mathrm{d}x) + \sum_{i=1}^k \xi_i^x \delta_{X_i^*}(\mathrm{d}x) + \sum_{j=1}^c \xi_j^y P_{Y_j^*}(\mathrm{d}x)$

The algorithm is straightforward, but relies on the computation of the weights at point (a): this is not optimal, since in general the explicit evaluation can be demanding. Nonetheless, Theorem 1 and Corollary 2 show that, conditionally on a suitable set of latent variables, the posterior representation simplifies greatly. Indeed, given $((X_i)_{i=1}^n, (Y_j)_{j=1}^m, U_1, U_2, p)$, the predictive distribution of the first sample is

$$\begin{aligned} m(\mathrm{d}x) \propto \theta\tau_{1,0}(U_1,U_2)P_0(\mathrm{d}x) + \sum_{(i,j)\in\Delta_p} \frac{\tau_{n_i+1,m_j}(U_1,U_2)}{\tau_{n_i,m_j}(U_1,U_2)}\delta_{X_i^*}(\mathrm{d}x) \\ + \sum_{(i,j)\in\Delta_p^1} \frac{\tau_{n_i+1,0}(U_1,U_2)}{\tau_{n_i,0}(U_1,U_2)}\delta_{X_i^*}(\mathrm{d}x) + \sum_{(i,j)\in\Delta_p^2} \frac{\tau_{1,m_j}(U_1,U_2)}{\tau_{0,m_j}(U_1,U_2)}P_{Y_j^*}(\mathrm{d}x). \end{aligned} \tag{13}$$

Those new weights, whose derivation can be found in Section S1.4, are easier to compute, as the next example shows.

Example 8 (Inverse Gaussian n-FuRBI with equal jumps)**.** For this case we obtain $\tau_{n,m}(u_1,u_2) = \int_{\mathbb{R}} s^{n+m}e^{-(u_1+u_2)s}\rho(\mathrm{d}s) := \tau_{n+m}(u_1+u_2)$, where $\rho(\mathrm{d}s)$ is the common marginal jump intensity. If the Lévy intensity is $\nu(\mathrm{d}s,\mathrm{d}x) = e^{-s/2}/(s^{3/2}\sqrt{2\pi})\mathrm{d}s\,\alpha(\mathrm{d}x)$ the resulting normalized CRM corresponds to the normalized inverse Gaussian process introduced in Lijoi et al. (2005). We then obtain $\tau_j(u) = 2^{j-1}\Gamma(j-1/2)/(\sqrt{\pi}(2u+1)^{j-1/2})$, where $u = u_1+u_2$. Thus, conditionally on the latent variables, we have

$$\begin{aligned} m(\mathrm{d}x) \propto \theta P_0(\mathrm{d}x) + \frac{2}{\sqrt{2U+1}} \sum_{(i,j)\in\Delta_p} \left(n_i + m_j - \frac{1}{2}\right)\delta_{X_i^*}(\mathrm{d}x) \\ + \frac{2}{\sqrt{2U+1}} \sum_{(i,j)\in\Delta_p^1} \left(n_i - \frac{1}{2}\right)\delta_{X_i^*}(\mathrm{d}x) + \frac{2}{\sqrt{2U+1}} \sum_{(i,j)\in\Delta_p^2} \left(m_j - \frac{1}{2}\right)P_{Y_j^*}(\mathrm{d}x), \end{aligned}$$

where $U = U_1 + U_2$. Sampling from this mixture is straightforward.

Thus we can derive a second marginal algorithm.

Marginal algorithm - 2

(a) Draw (U_1, U_2, p) from their conditional distributions specified in Section 5

(b) Draw X_{n+1} from $m(\mathrm{d}x)$ in (13)

However, even the full conditional distribution of p may not always be available in closed form, and it may be computationally intensive to evaluate, since it may have a very large support. When this is the case, we may encode the latent clustering structure in a more convenient way introducing two arrays of latent variables $\mathcal{C}_x = (c_{i,x})_{i \geq 1}$ and $\mathcal{C}_y = (c_{j,y})_{j \geq 1}$ such that $c_{i,x} = c_{i',x}$ denotes a tie between X_i and $X_{i'}$, $c_{j,y} = c_{j',y}$ denotes a tie between Y_j and $Y_{j'}$, while $c_{i,x} = c_{j,y}$ denotes a hyper-tie between X_i and Y_j. Moreover, we reorder the unique values in $\underline{X}^*_n$ and $\underline{Y}^*_m$, so that $X^*_c = X_i$ if and only if $c_{i,x} = c$ and $Y^*_c = Y_j$ if and only if $c_{j,y} = c$. Therefore, $\mathbb{P}[c_{n+1,x} = c \mid \mathcal{C}_x, \mathcal{C}_y, \underline{X}^*_n, \underline{Y}^*_m]$ is

$$\begin{cases} \mathbb{P}[X_{n+1} = X^*_c \mid \mathcal{C}_x, \mathcal{C}_y, \underline{X}^*_n, \underline{Y}^*_m], & \text{for } c \in \mathcal{C}_x \\ \displaystyle\int \mathbb{P}[X_{n+1} = x \mid \mathcal{C}_y, \underline{Y}^*_m]\, p_{Y^*_c}(x)\mathrm{d}x, & \text{for } c \in \mathcal{C}_y \setminus \mathcal{C}_x \\ \displaystyle\int \mathbb{P}[X_{n+1} = x]\, p_0(x)\mathrm{d}x, & \text{otherwise} \end{cases}$$

Finally, the distribution of p, given $\mathcal{C}_x$ and $\mathcal{C}_y$, is degenerate. Moreover, the posterior distribution of (U_1, U_2) given p is equal to the posterior distribution of (U_1, U_2) given $\mathcal{C}_x$ and $\mathcal{C}_y$. Therefore, we may build a marginal algorithm sampling $\mathcal{C}_x$ and $\mathcal{C}_y$ instead of p, without modifying the full conditional distribution for U_1 and U_2. The final marginal algorithm boils down to

Marginal algorithm - 3

(a) Draw (U_1, U_2) and $c_{n+1,x}$

(b) Sample X_{n+1} from $m(\mathrm{d}x) = \begin{cases} \delta_{X^*_{c_{n+1,x}}}(\mathrm{d}x), & \text{if } c_{n+1,x} \in \mathcal{C}_x \\ P_{Y^*_{c_{n+1,x}}}(\mathrm{d}x), & \text{if } c_{n+1,x} \in \mathcal{C}_y \setminus \mathcal{C}_x \\ P_0(\mathrm{d}x), & \text{otherwise} \end{cases}$

The advantage of such approach is twofold. First, we do not need to sample directly the full conditional distribution of p. Second, when the algorithm is applied to mixture models, as in section 6, sampling the unique values, instead of single observations, improves the mixing of the algorithm (cfr. Neal, 2000).

S4.2 Conditional posterior sampler based on the law of the CRV

To develop a conditional algorithm, we can sample from the distribution of $(\tilde{\mu}_1, \tilde{\mu}_2)$ and then normalize each draw to get an approximate realization of the random probabilities. Here we develop a general conditional sampler based on this approach that can be tailored to specific choices of the intensity in the prior.

By Theorem 1, we know that a posteriori $\mu = (\mu_1, \mu_2)$ is the sum of two components, that we call μ_{obs} and $\hat{\mu}$ and are such that

$$\mu_{obs} = \sum_{(i,j)\in\Delta_p} J_{i,j}\delta_{\left(X_i^*, Y_j^*\right)} + \sum_{(i,j)\in\Delta_p^1} J_{i,c+1}\delta_{\left(X_i^*, Z_i^x\right)} + \sum_{(i,j)\in\Delta_p^2} J_{k+1,j}\delta_{\left(Z_j^y, Y_j^*\right)}.$$

where $J_{i,j} = (J_{i,j}^1, J_{i,j}^2)$, and

$$\hat{\mu} = \left(\sum_{h=1}^{+\infty} S_h^1\delta_{(V_h,W_h)}, \sum_{h=1}^{+\infty} S_h^2\delta_{(V_h,W_h)}\right)$$

is a CRV with Lévy intensity $e^{-U_1 s_1 - U_2 s_2}\rho(\mathrm{d}s_1, \mathrm{d}s_2)G_0(\mathrm{d}x)$. Denote the marginal and joint tail integrals of $\hat{\mu}$ as

$$N_1(s) = \int_s^{+\infty}\int_0^{+\infty} e^{-U_1 s_1 - U_2 s_2}\rho(\mathrm{d}u_1, \mathrm{d}u_2), \quad N_2(s) = \int_0^{+\infty}\int_s^{+\infty} e^{-U_1 s_1 - U_2 s_2}\rho(\mathrm{d}u_1, \mathrm{d}u_2)$$

and

$$N(s_1, s_2) = \int_{s_1}^{+\infty}\int_{s_2}^{+\infty} e^{-U_1 s_1 - U_2 s_2}\rho(\mathrm{d}u_1, \mathrm{d}u_2).$$

Lastly, define the correspondent Lévy copula as $F(x, y) = N(N_1^{-1}(x), N_2^{-1}(y))$. If $F(x, y)$ is continuous on $[0, +\infty]^2$, the iterative conditional sampler based on the Ferguson and Klass algorithm (Ferguson and Klass, 1972) reads

(a) Generate μ_{obs} as follows

  (a1) Generate $(U_1, U_2, \mathbf{p})$ from the distributions specified in Section 5;

  (a2) Generate $J_{i,j} = (J_{i,j}^1, J_{i,j}^2)$ from the distributions specified in Theorem 1;

  (a3) Generate Z_i^x and Z_j^y from the distributions specified in Section 5.

(b) Generate an approximation of $\hat{\mu}$, given by $\left(\sum_{h=1}^{M} S_h^1\delta_{(V_h,W_h)}, \sum_{h=1}^{M} S_h^2\delta_{(V_h,W_h)}\right)$ as follows

  (b1) Generate $\xi_1^x, \ldots, \xi_M^x$ from a Poisson Process with unit rate;

(b2) Generate $\xi_1^y, \ldots, \xi_M^y$ from $\xi_h^y \sim \frac{\partial}{\partial x} F(x, \xi) \Big|_{x=\xi_h^x}$

(b3) Determine (S_h^1, S_h^2) solving

$$\xi_h^x = N_1(S_h^1) \qquad \xi_h^y = N_2(S_h^2)$$

(b4) Generate (V_h, W_h) from G_0.

(c) Obtain a draw from $\tilde{p}_1$ as follows

$$\tilde{p}_1 \approx \frac{\sum_{h=1}^{M} S_h^1 \delta_{V_h} + \sum_{(i,j)\in\Delta_p} J_{i,j}^1 \delta_{X_i^*} + \sum_{(i,j)\in\Delta_p^1} J_{i,c+1}^1 \delta_{X_i^*} + \sum_{(i,j)\in\Delta_p^2} J_{k+1,j}^1 \delta_{Z_j^y}}{\sum_{h=1}^{M} S_h^1 + \sum_{(i,j)\in\Delta_p} J_{i,j}^1 + \sum_{(i,j)\in\Delta_p^1} J_{i,c+1}^1 + \sum_{(i,j)\in\Delta_p^2} J_{k+1,j}^1}.$$

An analogous approximation can be computed for $\tilde{p}_2$.

S4.3 Conditional posterior sampler for gamma process with equal jumps

Alternatively, a second strategy for conditional algorithms is to sample approximate draws from the posterior distribution of the random probabilities $(\tilde{p}_1, \tilde{p}_2)$. We provide an example for gamma FuRBI CRMs with equal jumps.

In the case of a process with equal jumps, we know from the definition that the measures in the product space are $p_1 = p_2 = p$. Therefore, posterior inference can be conducted without loss of generality on

$$p = \sum_{k\geq 1} \bar{W}_k \delta_{(\theta_k, \phi_k)}, \quad \text{with } (\theta_k, \phi_k) \overset{\text{i.i.d.}}{\sim} G_0(\cdot),$$

where $\{\bar{W}_k\}_k$ are the weights of a Dirichlet process, which can defined through the popular stick-breaking construction (Sethuraman, 1994). In this context, Ishwaran and James (2001) developed a conditional algorithm for hierarchical mixture models, called *blocked Gibbs* sampler, based on the approximation

$$p \approx \sum_{k=1}^{N} \bar{W}_k \delta_{(\theta_k, \phi_k)}, \quad \text{for large } N.$$

Exploiting the appealing analytical properties of the Dirichlet process, it is possible to devise simple formulae for the posterior distribution of the N jumps and N locations: see Section 5 of Ishwaran and James (2001) for more details.

S4.4 Sampling from mixture models using marginal algorithms

Consider the mixture model defined in Section 6.1. Starting from Algorithm 2 in Section $S4.1$, we devise a Gibbs sampler for drawing from the posterior distribution of $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$.

Denoting by $\boldsymbol{X}^t = (X_1^t, \ldots, X_n^t)$ and $\boldsymbol{Y}^t = (Y_1^t, \ldots, Y_n^t)$ the vectors sampled at step t, the algorithm reads

1. Initialize at random $\boldsymbol{X}^0$ and $\boldsymbol{Y}^0$.
2. For any $t \geq 1$ do:
   (b.1) Draw $(U_1, U_2, \mathbf{p})$ given $\boldsymbol{X}^{t-1}$ and $\boldsymbol{Y}^{t-1}$, from the distributions specified in Theorem 1.
   (b.2) Draw $\boldsymbol{X}_n$, given $(U_1, U_2, \mathbf{p})$ as follows: for any i sample X_i^t from

$$q(\mathrm{d}x \mid \boldsymbol{X}_{-i}^t) = q_{i,0}(U_1, U_2)P_0(\mathrm{d}x) + \sum_{(i,j)\in\Delta_{\boldsymbol{p}}} q_{i,j}(U_1, U_2)\delta_{X_i^*} + \sum_{(i,j)\in\Delta_{\boldsymbol{p}}^1} q_{i,j}^1(U_1, U_2)\delta_{X_i^*}(\mathrm{d}x) + \sum_{(i,j)\in\Delta_{\boldsymbol{p}}} q_{i,j}^2(U_1, U_2)P_{Y_j^*}(\mathrm{d}x),$$

   where $\boldsymbol{X}_{-i}^t = \left(X_1^t, \ldots, X_{i-1}^t, X_{i+1}^{t-1}, \ldots X_n^{t-1}\right)$, with unique values $(X_1^*, \ldots, X_k^*)$ and multiplicities $(n_1, \ldots, n_k)$. Analogously, $(Y_1^*, \ldots, Y_c^*)$ denotes the unique values in $\boldsymbol{Y}^{t-1}$ with multiplicities $(m_1, \ldots, m_c)$. The mixing proportions are given by

$$\begin{aligned}
q_{i,0}(U_1, U_2) &\propto \theta\tau_{1,0}(U_1, U_2)\int_{\mathbb{X}} f(W_i \mid x)P_0(\mathrm{d}x),\\
q_{i,j}(U_1, U_2) &\propto \frac{\tau_{n_i+1,m_j}(U_1, U_2)}{\tau_{n_i,m_j}(U_1, U_2)} f(W_i \mid X_i^*),\\
q_{i,j}^1(U_1, U_2) &\propto \frac{\tau_{n_i+1,0}(U_1, U_2)}{\tau_{n_i,0}(U_1, U_2)} f(W_i \mid X_i^*),\\
q_{i,j}^2(U_1, U_2) &\propto \frac{\tau_{1,m_j}(U_1, U_2)}{\tau_{0,m_j}(U_1, U_2)}\int_{\mathbb{X}} f(W_i \mid x)P_{Y_j^*}(\mathrm{d}x)
\end{aligned}$$

   (c) Sample $\boldsymbol{Y}^t$ similarly to point (b).

Once a sample of $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$ is available, sampling new observations X_{n+1} and Y_{n+1} proceeds as explained in Section $S3.1$.

S5 Additional simulation studies

S5.1 Additional simulation scenarios

We consider the same setting of Section 6.2 in the main manuscript, with different data generating distributions. Formally we have

$$W_i \overset{\text{i.i.d.}}{\sim} p(\cdot - 10), \quad \text{var}_j \overset{\text{i.i.d.}}{\sim} p(\cdot - v),$$

where $v \in [-16, 16]$ and $p(\cdot)$ is the density function of a zero mean random variable. In the main manuscript we let $p(\cdot) = N(\cdot \mid 0, 1)$, while here we consider three different choices

$$p_1(\cdot) = \text{Exp}(\cdot \mid 1), \quad p_2(\cdot) = 0.5N(\cdot \mid 5, 1) + 0.5N(\cdot \mid -5, 1), \quad p_3(\cdot) = t(\cdot \mid 3),$$

where $t(\cdot \mid q)$ denotes the density of a Student's t distribution with q degrees of freedom. We let $i = 1, \ldots, 20$, $j = 1, \ldots, 100$ and consider the same nonparametric models of Section 6.2, with Gaussian kernel. Therefore, the prior specification is misspecified in the first and third case, with different tail behaviours of the kernel with respect to the true data generating mechanism. This implies a more complex behaviour of the latent clustering structure: indeed the posterior distribution places positive mass to more than one clusters, in order to accommodate for the misspecification. The mean integrated error for the three cases is depicted in Figure S2, for different values of v. The interpretation is similar to the one discussed in Section 6.2: the FuRBI specification yields an advantage especially when v is far from 0, corresponding to the prior mean, and from 10, when the means of the two groups coincide. Indeed, in the first case the borrowing provides little information, while in the second one exchangeability holds.

The second setting, corresponding to the two-components mixture, apparently seems more problematic for the FuRBI model, which yields a less distinct advantage. Clearly, when v is close to zero the exchangeable and hierarchical models are favoured, since the two true distributions share one of the modes. Moreover, the availability of only 20 observations for the first group makes it more difficult to both detect the presence of two clusters and tune appropriately the correlation. Indeed, the left part of figure S3 depicts the error when 50 observations for the first group are collected: as expected, the performances of the FuRBI approach significantly improve.

Finally, the right part of figure S3 shows the error when the two distributions are different: the first group is endowed with a Student's t density, while the second one is exponentially distributed. Notice that the two groups are now very far in distributional sense, especially in terms of tail behaviour. The plot indicates an interesting trade-off: when v is far from the prior mean (i.e. 0) the FuRBI approach allows to alleviate the prior misspecification, otherwise borrowing information from very different distributions may be detrimental.

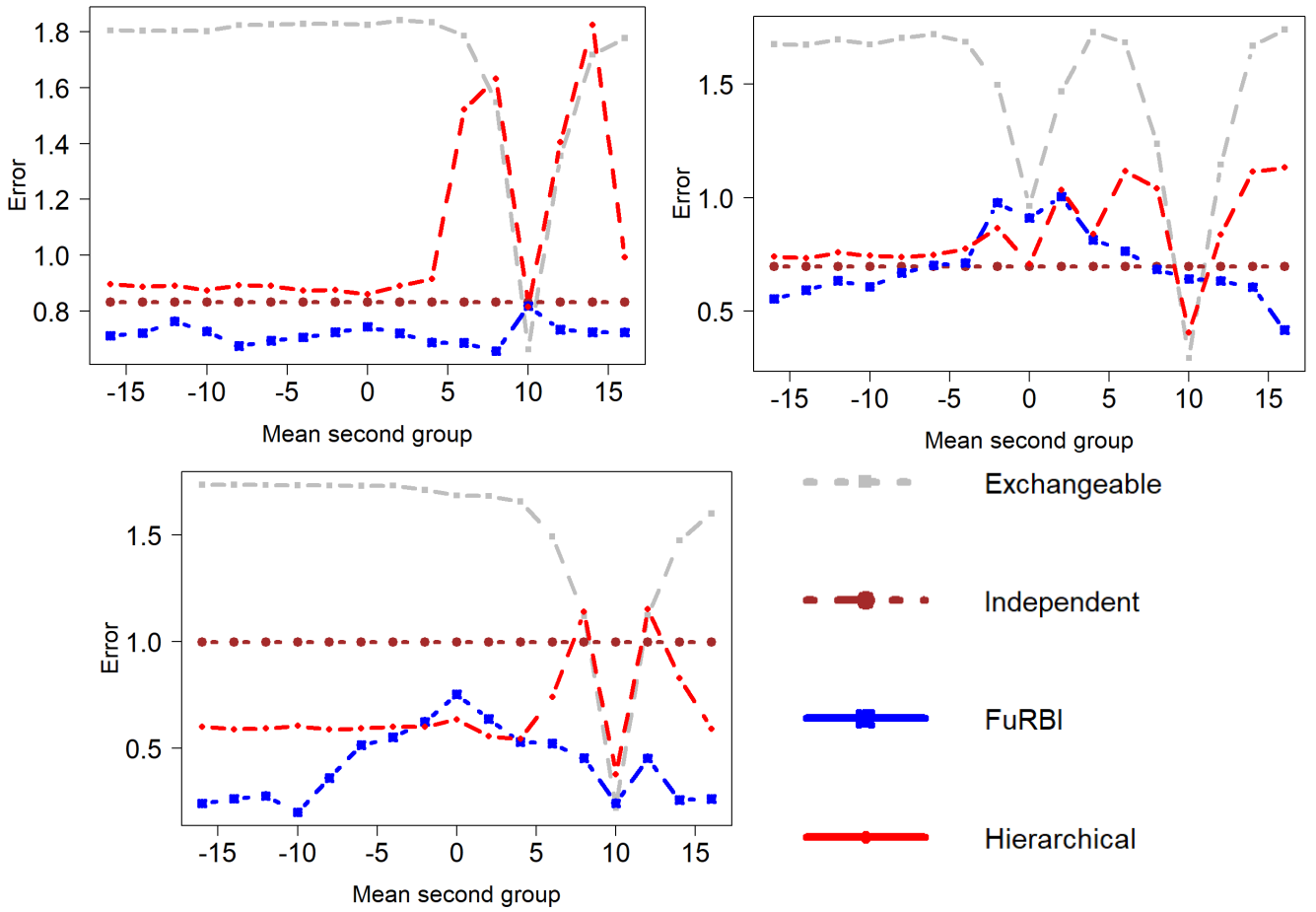


Figure S2: Mean integrated error (computed on a grid and as the median over 50 different samples) for the four models, as the true mean of the second group varies. Rotating clockwise from the top left panel: data generated from shifted exponential, mixtures of two Gaussians and shifted Student's t distributions.

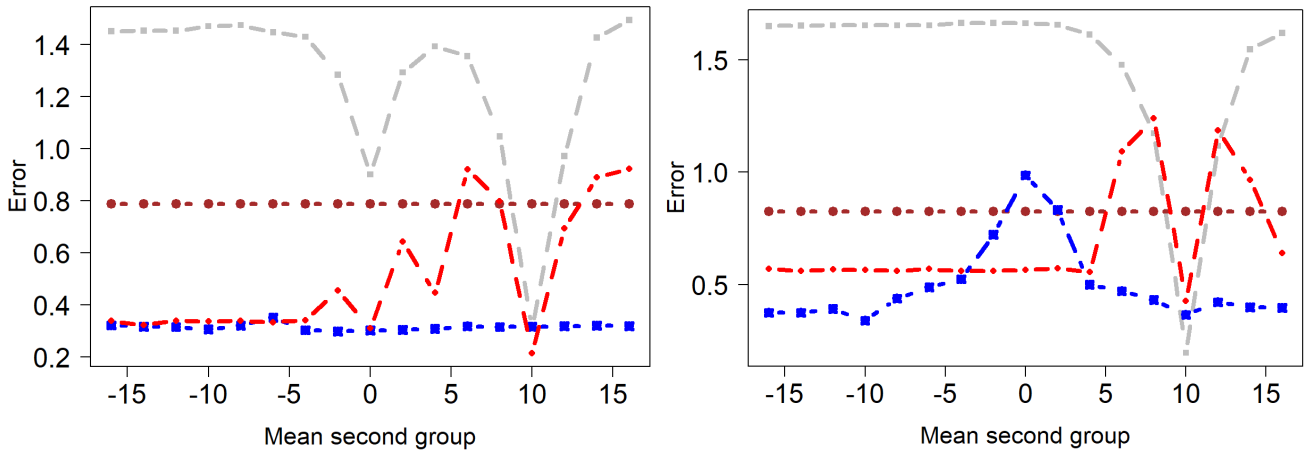


Figure S3: Mean integrated error (computed on a grid and as the median over 50 different samples) for the four models, as the true mean of the second group varies. Left: data generated from mixtures of two Gaussians (50 observations for the first group). Right: data generated from shifted Student's t (first group) and shifted exponential (second group) distributions.

S5.2 Logit stick-breaking prior and borrowing of information

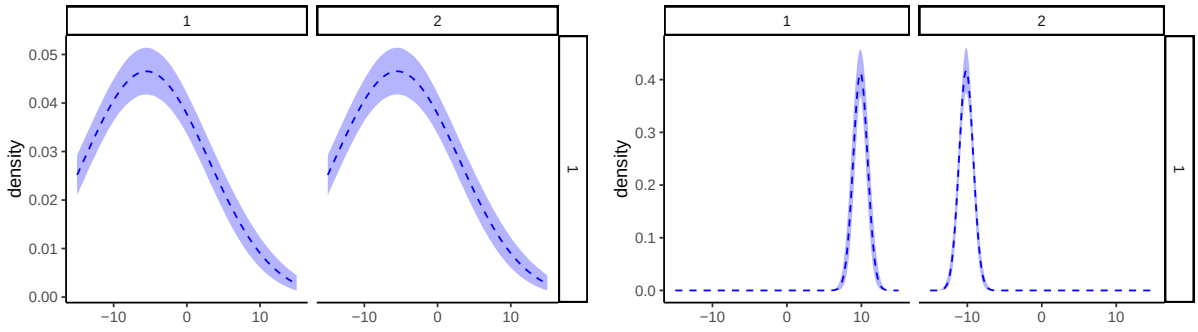
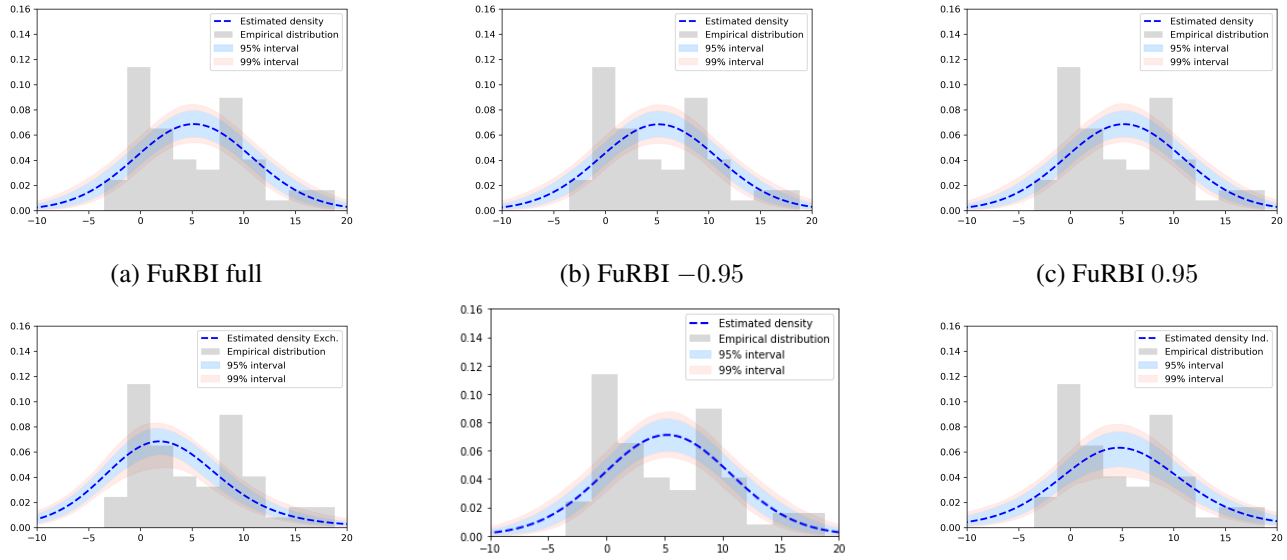


Figure S4: Left panel: density estimates for the logit stick-breaking model with only dependent weights, and thus, $\rho_0 = 1$. Right panel: density estimates for the logit stick-breaking model with dependent weights and atoms. Shaded areas denote 95% credible intervals. Data are simulated according to $W_i \overset{\text{i.i.d.}}{\sim} N(\cdot \mid 10, 1)$, for $i = 1, \ldots, 20$ (for sample n.1), and $\text{var}_j \overset{\text{i.i.d.}}{\sim} N(\cdot \mid -10, 1)$, for $j = 1, \ldots, 100$ (for sample n.2).

Figure S4 is based on the same data of Section 6.2. See Rigon and Durante (2021) for the model and the associated algorithm.

S6 Predicting stocks and bonds returns: additional results

S6.1 Density estimation for bond returns



(a) FuRBI full (b) FuRBI -0.95 (c) FuRBI 0.95

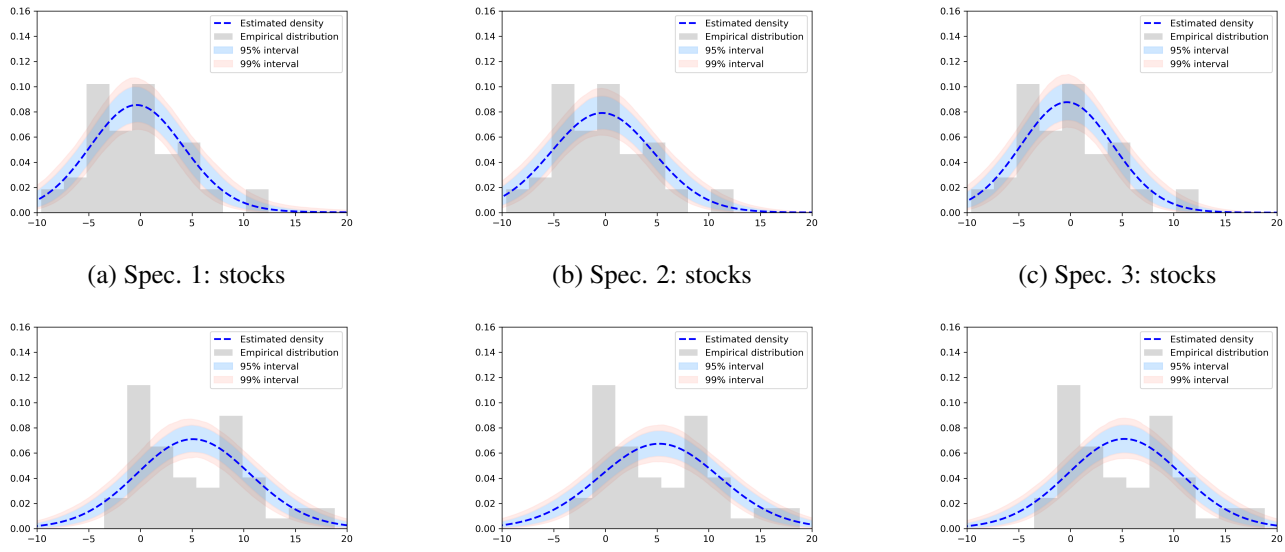
(d) Exchangeable model (e) GM-dependent model (f) Independent model

Figure S5: Density estimates for bonds returns.

S6.2 Sensitivity analysis

Figure S6 shows the results obtained with different specifications of the hyperparameters, which are

- Specification n.1: $\lambda_j = 0.1$, $\alpha_j = 3$, and $\beta_j = 3$, $j = 1, 2$,
- Specification n.2: $\lambda_j = 0.1$, $\alpha_j = 1.5$, and $\beta_j = 4.5$, $j = 1, 2$,
- Specification n.3: $\lambda_j = 0.01$, $\alpha_j = 0.1$, and $\beta_j = 0.2$, $j = 1, 2$.



(a) Spec. 1: stocks (b) Spec. 2: stocks (c) Spec. 3: stocks

(d) Spec. 1: bonds (e) Spec. 2: bonds (f) Spec. 3: bonds

Figure S6: Sensitivity analysis: density estimates for bonds returns.

S6.3 Posterior distribution of ρ_0

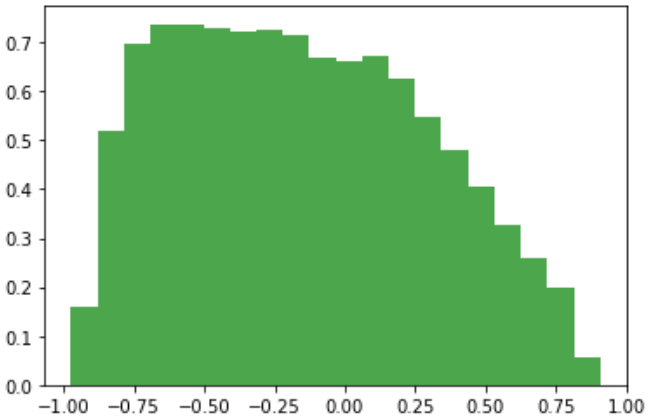


Figure S7: Posterior distribution of ρ_0 for the analysis in Section 6.3.

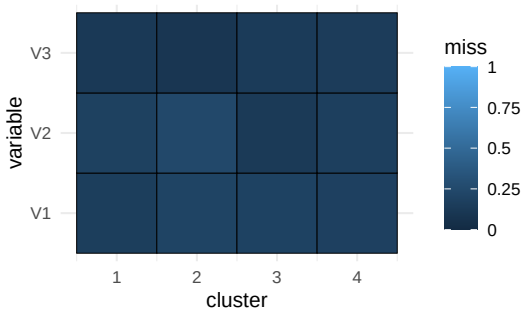
S7 Clustering multivariate data with missing entries: additional details

S7.1 Choosing the hyperparameters

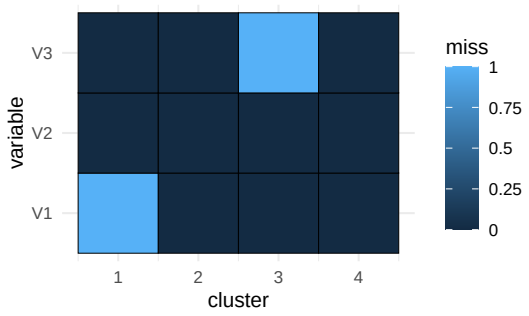
Assume $P = 3$, as in the simulation study of Section 6.4: the general case follows accordingly. In this case $I = \{\emptyset, (1), (2), (3), (1,2), (2,3), (1,3), (1,2,3)\}$. In order to specify the prior, assumptions on the missing generating mechanism should be made. The missing completely at random (MCAR) assumption implies that each observation $W_i^{(x)}$, for $x \in I$, is the result of randomly eliminating entries from an (unobserved) complete observation W_i. For instance, $W_i^{(1)} = (w_{2,i}, w_{3,i})$ is obtained from a latent $W_i = (w_{1,i}, w_{2,i}, w_{3,i})$ after eliminating the first entry. Under this assumption the latent complete observations W_i are exchangeable, because the original value of W_i is independent from the mechanism that generates the missing values. Thus, there exists $\tilde{q}$ such that $W_i \mid \tilde{q} \overset{iid}{\sim} \tilde{q}$ and $\tilde{q}_x$ is the projection of $\tilde{q}$ onto coordinates different than x, e.g. $\tilde{q}_{(1)}(\cdot, \cdot) \overset{a.s}{=} \int \tilde{q}(\mathrm{d}x_1, \cdot, \cdot)$. This implies that the weights of $\tilde{q}_x$ should be almost surely the same for every x. Instead, if the missing mechanism is not completely at random, $\tilde{q}_x$ can not be described as the projection of a unique $\tilde{q}$. Indeed the missing mechanism may be informative, leading to sample-specific features. Therefore, the choice of an additive n-FuRBIs allows $\tilde{q}_x$ to have sample-specific components when needed.

As for the baseline distribution G_0 on $\boldsymbol{\mu}$, suppose that an hyper-tie is sampled between an observation $(w_{2,i}, w_{3,i})$ from sample "(1)" and one observation $(w_{1,i}, w_{3,i})$ from sample "(2)", thus assigning the two observations to the same cluster. G_0 is then used to sample the corresponding locations: (X_2^*, X_3^*) and (Y_1^*, Y_3^*). Since we want to interpret the hyper-tie between incomplete observations as a tie between complete observations, we must have $X_3^* = Y_3^*$, while X_2^* and Y_1^* are sampled jointly with a certain correlation $\rho_{1,2}$ and depending on X_3^* through correlations $\rho_{1,3}$ and $\rho_{2,3}$. Therefore, since coordinates corresponding to the same original variable should be assigned the same value, G_0 is actually degenerate on a $P = 3$ dimensional space. In the simulation and real data application G_0 is a 3-variate normal, whose correlation matrix ρ_0 depends on correlation parameters ρ_{12}, ρ_{23}, ρ_{13} on which a truncated uniform hyperprior is used, where the truncation ensures that the matrix is almost-surely positive-definite. Since the data are centered, the mean of G_0 is instead fixed equal to a vector of all 0. Moreover, an independent Gamma$(3, 3)$ prior is assigned to the three variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$. Finally, the concentration parameter θ is set equal to 0.1 in order to favor sparsity, i.e., a lower number of clusters.

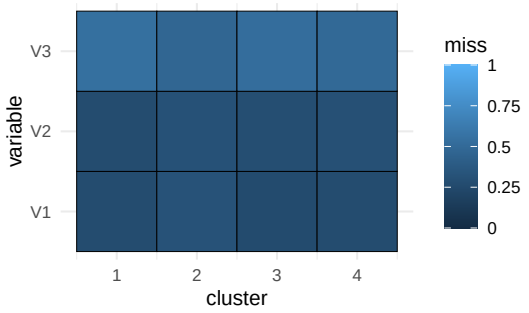
S7.2 Simulating scenarios: missing data distribution



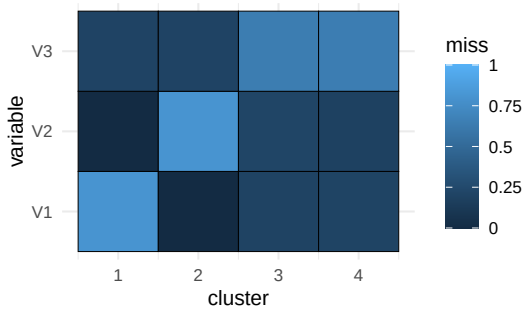
(a) MCAR 16.1% missing entries



(b) MNAR 17.7% missing entries



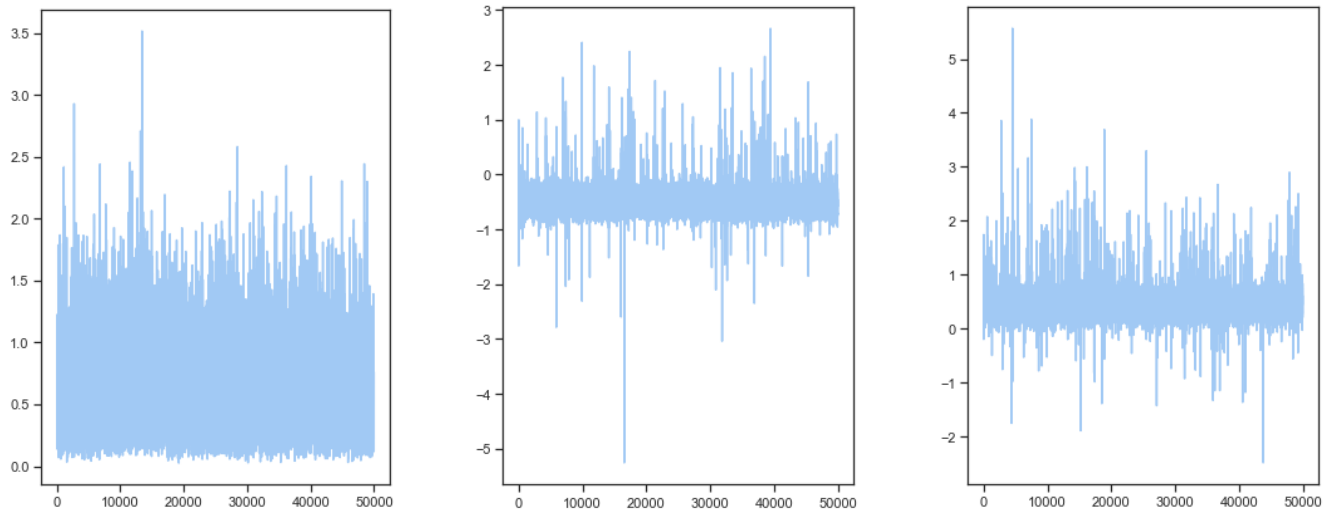
(c) MCAR 35.9% missing entries



(d) MNAR 34% complete observations

Figure S8: Percentages of missing entries of each variable-cluster pair.

S8 Mixing performance of the MCMC chains



(a) Concentration parameter θ. (b) Stock n.1 cluster location. (c) Bond n.1 cluster location.

Figure S9: Trace plots of the MCMC chain used in the real data analysis of Section 6.3.

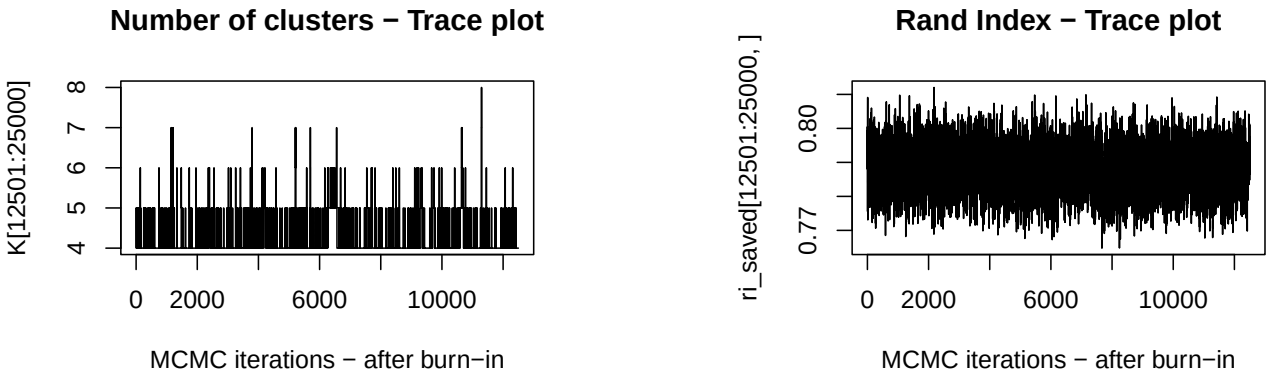


Figure S10: Trace plots of the MCMC chain of simulation study n.1 in Section 6.4, for the additive n-FuRBI model with $z = 0.5$. Left: number of clusters. Right: Rand index.

Model	ESS / N Rand index	ESS / N num. clusters
Additive n-FuRBIs, z = 0.2	0.1957	0.0518
Additive n-FuRBIs, z = 0.5	0.1994	0.0413
Additive n-FuRBIs, z = 0.8	0.1253	0.0596
DPM	0.1623	0.0227

Table 5: Effective Sample Size (ESS) per iteration in simulation study n.1 of Section 6.4 with $1,000$ observations.

	total sample size	dimension of data point	Type of algorithm	code language	average time per iter (in sec)
Financial data - Sec. 6.3	$n = 104$	1	marginal	Python	0.12
Simulation studies - Sec. 6.4	$n = 1000$	3	marginal	R	2.41
Brandsma data - Sec. 6.4	$n = 4106$	4	marginal	R	8.75

Table 6: Computational time in second per one iteration of the MCMC chain with n-FuRBIs. Codes are run on an Intel Xeon W-1250 processor. Note that the in the last two lines not only the sample size is higher but also the data are multivariate.