BAYESIAN NONPARAMETRIC MIXTURE MODELING FOR TEMPORAL DYNAMICS OF GENDER STEREOTYPES

BY MARIA DE IORIO, STEFANO FAVARO*, ALESSANDRA GUGLIELMI AND LIFENG YE

*Yong Loo Lin School of Medicine, NUS; Singapore Institute for Clinical Sciences, A*STAR ; University College London*
mdi@nus.edu.sg

Università di Torino and Collegio Carlo Alberto
stefano.favaro@unito.it

Politecnico di Milano
alessandra.guglielmi@polimi.it

University College London
lifeng.ye.13@ucl.ac.uk

The study of temporal dynamics of gender and ethnic stereotypes is an important topic in many disciplines at the intersection between statistics and social sciences. In this paper, we make use of word embeddings, a common tool in natural language processing, and of Bayesian nonparametric mixture modeling for the analysis of temporal dynamics of gender stereotypes in adjectives and occupation over the 20th and 21st centuries in the United States. Our Bayesian nonparametric approach relies on a novel dependent Dirichlet process prior, and it allows for both dynamic density estimation and dynamic clustering of adjective embedding and occupation embedding biases in a hierarchical setting. Posterior inference is performed through a particle Markov chain Monte Carlo algorithm which is simple and computationally efficient. An application to time-dependent data for adjective embedding bias and for occupation embedding bias shows that our approach enables the quantification of historical trends of gender stereotypes, and hence allows to identify how specific adjectives and occupations have become more closely associated with a female rather than male over time.

1. Introduction. The study of the changes over time in gender and ethnic stereotypes is an important topic in many disciplines, as it lies at the intersection between quantitative and social sciences (Williams and Best, 1990; Basow, 1992; Holmes and Meyerhoff, 2008; Coates, 2016). In recent years, machine learning methods have become increasingly popular to analyse stereotypes dynamics, with the goal of capturing the dynamic formation, maintenance, and transformation of stereotypes. Moreover, such methods have proved to be crucial in reducing time-consuming and expensive manual analysis and in scaling across types of stereotypes, time periods, and languages. In particular, natural language processing techniques have been often applied to measure, quantify, and compare gender and ethnic stereotypes over time. To this end, a common tool is provided by word embeddings, a learned word representation for text analysis (typically in the form of a real-valued vector) such that words with similar meaning have a similar representation. Our starting point is the work of Garg et al. (2018) who exploits word embeddings to study historical trends in gender and ethnic stereotypes, demonstrating how the temporal dynamics of the embeddings are able to capture actual changes in stereotypes and attitudes towards women and ethnic minorities in the 20th

*Also affiliated to IMATI-CNR "Enrico Magenes" (Milan, Italy).

Keywords and phrases: Autoregressive models, Bayesian nonparametrics, dependent Dirichlet processes, dynamic density estimation and clustering, gender stereotypes, mixture modeling, particle Markov chain Monte Carlo, word embeddings

and 21st centuries in the United States. More in details, a word embedding represents each word as a high-dimensional real-valued vector, whose geometry captures local and global semantic relations between words, e.g., words with a representation closer together in a vector space correspond to more similar words (Collobert et al., 2011). These models are typically trained automatically on large corpora of text, such as collections of Google News articles or Wikipedia, and are known to capture relationships not found through simple co-occurrence analysis. As stereotypes are likely to be present, even if subtly, in large corpora of training texts, word embeddings have been successfully employed to capture them. (Bolukbasi et al., 2016; Caliskan, Bryson and Narayanan, 2017; Zhao et al., 2017). The goal of this work is to investigate the clustering dynamics of gender stereotypes as captured by adjectives and occupation-related words over the 20th and 21st centuries in the United States. We exploit the potential of word embeddings combined with flexible tools from Bayesian nonparametric (BNP) mixture modeling to model time-dependent data on bias towards women as described by *adjective* and *occupation* embeddings. From a methodological perspective, BNP dynamic mixtures require the specification of a prior process for a collection of random distributions indexed by time, whose realizations are dependent. These models are usually built as extension of the Dirichlet process (DP) prior (Ferguson, 1973) and exploit the discreteness of the DP prior, allowing for flexible and robust estimation of time-evolving densities as well as for dynamic clustering of items (see, for instance, work by Taddy (2010) and DeYoreo and Kottas (2018) focusing on density estimation). A BNP approach is particularly appealing for our applications, since a preliminary analysis shows that a parametric dependence structure is unable to fully capture the data complexity and there is no information on the number of components, if a finite mixture approach were to be adopted. As such data-driven clustering methods are preferred, able to accommodate heterogeneity, overdispersion and outliers. From a computational perspective, mixture models based on DP priors offer many advantages and a plethora of exact and approximate computational algorithms for posterior inference is available.

1.1. *Our contributions.* We consider word embeddings trained on Corpus of Historical American English (COHA) (Hamilton, Leskovec and Jurafsky, 2016) for eleven decades between year 1900 and 2000, together with the list of adjective and occupation words provided by Garg et al. (2018). We are interested in the analysis of data on (standardized) adjective and occupation embedding biases towards women, for each word in the corresponding list. Here by adjective embedding bias, we mean a bias representation obtained by the embedding of an adjective. Similarly, occupation bias is obtained from the embedding of a occupation-related word. A negative value of the bias implies that the embedding more closely associates the adjective or the occupation-related word with men, because the distance between the words is closer to men than women. We refer to a negative value of the embedding bias as "bias against women". Note that gender bias corresponds to either negative or positive values of the embedding bias. We develop a novel BNP dynamic mixture model for time-dependent data for adjective and occupation embedding bias, able to quantify changes in gender stereotypes over the 20th and 21st centuries in the United States. Clustering, i.e. finding homogeneous subgroups, is critical in gender stereotyping. For example, in Six and Eckes (1991) the clustering structure is essential to uncover cognitive ordering principles underlying gender stereotypes. Lewis et al. (2020) use word-embeddings to quantify the presence of gender stereotypes in $200,000$-word corpus of 247 books for children, obtaining 100 clusters of words through k-means clustering based on their coordinates. Each cluster is interpreted ex-post as female-biased (containing words such as kisses, loved, care, soup, eggs, milk, pie), neutral or male-biased clusters (e.g. axe, blade, knife, car, bicycle, trains, policemen, guard, sailor) according to the mean-rated genderedness of the words in the cluster as compared with the mean-rated genderedness of all words in the sample.

Here we propose a BNP dynamic mixture model that relies on a novel dependent DP prior for a sequence of discrete random probability measures $(G_t)_{t\geq 1}$, with t indexing discrete time. Similarly to other dependent DPs, we exploit the stick-breaking construction of the DP (Sethuraman, 1994) to introduce dependency among the G_t's. To this end, we employ a suitable copula-based transformation of a Gaussian autoregressive process of order 1 (Guolo and Varin, 2014). The resulting sequence of random probability measures $(G_t)_{t\geq 1}$ belongs to the broad class of dependent DP priors of MacEachern (2000), such that: i) $(G_t)_{t\geq 1}$ has an autoregressive structure of order 1; ii) G_t is a DP, for any $t \geq 1$. Therefore, the law of $(G_t)_{t\geq 1}$ provides an autoregressive order 1 DP (AR1-DP) nonparametric prior. We apply the AR1-DP prior to define a BNP mixture model for time-dependent data of biases $\{(Y_{t1}, \ldots, Y_{tn})\}_{t=1,\ldots,T}$, where n is the number of embedding biases. We assume that the Y_{tj}'s are modeled as

$$
\begin{aligned}
Y_{tj} \mid \theta_{tj} &\overset{\text{ind}}{\sim} k(\cdot\,; \theta_{tj}) \quad j = 1, \ldots, n \\
\theta_{t1}, \ldots, \theta_{tn} \mid G_t &\overset{\text{iid}}{\sim} G_t \quad t = 1, 2, \ldots, T \\
(G_t)_{t\geq 1} &\sim \text{AR1-DP}
\end{aligned} \tag{1}
$$

with $k(\cdot\,; \theta)$ being a density function parameterized by $\theta \in \Theta \subset \mathbb{R}^p$, e.g. a Gaussian density function with mean θ. The mixture model (1) may be viewed as a dynamic counterpart of the popular DP mixture model (Lo, 1984) and provides a flexible tool to describe the clustering dynamics of adjective and occupation embedding bias in a hierarchical setting . Indeed, the AR1-DP mixture model (1) allows cluster memberships to change over time, creating new clusters and removing existing ones. In particular, because of the autoregressive structure of order 1 of the AR1-DP prior, the clustering configuration at time $t + 1$ depends on the clustering configuration at time t. The AR1-DP covers a wide range of dependence structure, from independence across time periods to identical clustering structure over time. Model (1) allows for posterior inference at a specific time point to borrow strength from the clustering distribution at different times. Furthermore, it allows the inclusion of covariate information, when available. Posterior inference can be performed through a particle Markov chain Monte Carlo (MCMC) algorithm (Andrieu, Doucet and Holenstein, 2010) which is simple and computationally efficient.

We apply the AR1-DP mixture model to adjective embedding bias data (see Section 3.1) and to occupation embedding bias data (in Section 3.2), showing how it allows to quantify historical trends of gender bias, and also to identify how specific adjectives and occupations have become more closely associated with a specific gender over time. For occupation embedding bias data, posterior predictive densities (see Figure 8) confirm a preliminary descriptive statistical analysis (see the boxplots over time in Figure 12), showing that bias against women tends to mitigate from 1900 to 2000. That is, estimated densities in Figure 8 are explained by two mixture components from 1900 to 1970, and by one mixture component in 1980, 1990 and 2000. In the latter case, the estimated mode moves towards larger values of the support of the distribution from 1980 to 2000. The number of estimated clusters (see Section 3.1) equals two for decades $1900, \ldots, 1970$, while equals 1 in decades 1980, 1990, 2000. Similar results are obtained for adjective embedding bias data. For all decades, posterior predictive densities (Figure 10) show three well-separated peaks with different proportions; see also the boxplots of adjective embeddings in Figure 13. For ease of interpretation, estimated clusters are labelled as man/neutral/woman clusters, with the inclusion in one group corresponding to conventional stereotypes. Our approach identifies a different structure, both in density estimation and clustering, between the last three decades and the previous ones. This might be due to the well-known socio-economic changes occurred in the period 1960-1970,

which contributed to a “revolution" in the society and the spoken language (see, for instance, Boltanski and Chiapello, 2006). In particular, we refer to what in the literature is called as the second wave of feminism, a period of time between late 60s and early 70s, when in the US women's rights and women's liberation became mass movements (Nicholson, 2010; Hewitt, 2012). At the same time, there was a critical change in the language, with the emergence of a new psychological/sociological language which acknowledged the sex roles stereotypes (Altman, 2003). See Section 3.3 for more details.

1.2. *Related work.* There exists a vast literature on dependent random probability measures indexed by discrete time, e.g., Griffin and Steel (2006), Caron, Davy and Doucet (2007), Dunson, Pillai and Park (2007),Caron et al. (2008), Dunson and Park (2008), Rodriguez and ter Horst (2008), Taddy (2010), Rodriguez and Dunson (2011), Griffin and Steel (2011), Nieto-Barajas et al. (2012), Di Lucca et al. (2013), Bassetti, Casarin and Leisen (2014), Xiao, Kottas and Sansó (2015), Gutiérrez, Mena and Ruggiero (2016), DeYoreo and Kottas (2018). The prior processes developed by Taddy (2010) and DeYoreo and Kottas (2018) are closely related to the AR1-DP prior. In particular, DeYoreo and Kottas (2018) define a dependent prior through a transformation of a Gaussian autoregressive process of order 1 which, however, imposes more constraints on the dependence structure than the proposed copula-based transformation. An alternative approach to dynamic clustering is offered by the work of Page, Quintana and Dahl (2021), who introduce a prior process directly on a sequence of partitions indexed by discrete time.

Dependent DPs defined through copula-based stick-breaking representations are popular in the BNPs, and they were first suggested in MacEachern (2000). Rodríguez, Dunson and Gelfand (2010) introduce dependence across locations through a latent Gaussian copula model as the mechanism for selecting the atoms. Rodriguez and Dunson (2011) propose a more general AR-1stick-breaking process with weights constructed as probit transformations of normal random variables, whereas Pati, Dunson and Tokdar (2013) consider dependent mixtures of Gaussians which include probit stick-breaking mixtures of Gaussians. Arbel, Mengersen and Rousseau (2016) propose a copula-based covariate-dependent stick-breaking process for an environmental application featuring species-by-site count data, with a copula transformation similar to ours. Finally, we highlight that our prior process is a special case of the general framework developed by Barrientos, Jara and Quintana (2012), who provide an alternative definition of MacEachern's Dependent Dirichlet process. Their construction allows specifying finite dimensional distributions of stochastic processes with given marginal distributions, by exploiting the connection between stochastic processes and copulas.

1.3. *Organization of the paper.* The paper is structured as follows. In Section 2 we introduce and characterise the AR1-DP mixture model: i) we define the AR1-DP prior as a novel dependent DP; ii) we present a particle MCMC algorithm for posterior inference; iii) we compare the AR1-DP mixture model with the models of Taddy (2010) and DeYoreo and Kottas (2018). In Section 3 we apply the AR1-DP mixture model to time-dependent data on occupation embedding bias (Section 3.1) and adjective embedding bias (Section 3.2). We briefly discuss our findings in connection with sociological literature in Section 3.3. Section 4 concludes the paper with a discussion and directions for future research. In Appendix we provide the list of words used in the application as well as the definition of occupation and adjective bias. We also describe the two datasets and perform an exploratory analysis to investigate the presence of a temporal component in the clustering structure. The (online) Supplementary Material contains some details on the MCMC algorithm for posterior inference and, in particular, on the particle MCMC step, an extensive simulation study illustrating the performance of the AR1-DP mixture model and additional figures.

2. The AR1-DP mixture model. We present the AR1-DP mixture model (1) and compare it with competitor models in BNPs. We start by introducing the AR1-DP prior at the latent level of (1), and by devising a particle MCMC algorithm for posterior inference. Observe that, according to (1), we are assuming that time-dependent data of biases $\{(Y_{t1}, \ldots, Y_{tn})\}_{t=1,\ldots,T}$ for each t are conditionally i.i.d. according to the random density function

$$f_t(y) = \int k(y\,;\,\theta) G_t(d\theta), \tag{2}$$

with the AR1-DP prior being a prior distribution on $(G_t)_{t\geq 1}$. The AR1-DP prior provides a new instance of dependent DP priors (MacEachern, 2000), i.e. a prior distribution for a collection of dependent random probability measures indexed by discrete times. In this section, we present a constructive definition of the AR1-DP prior, which combines a copula-based transformation of a Gaussian autoregressive process of order 1 (Guolo and Varin, 2014) with the stick-breaking construction of the DP prior (Sethuraman, 1994). The constructive definition of the AR1-DP prior plays a critical role in our study of the temporal dynamics of gender stereotypes: i) it brings out the role of prior parameters, especially for tuning the degree of dependence between random probability measures at different time points; ii) it leads to a particle MCMC for posterior inference which is simple and computationally efficient.

2.1. *AR1-DP priors.* Let $\mathrm{N}(\mu, \sigma^2)$ denote a Gaussian distribution with mean μ and variance σ^2, let $\epsilon \sim \mathrm{N}(0,1)$ and let $\Phi(\cdot)$ denote the cumulative distribution function of ϵ. Recall that if $F(\cdot\,;\,a,b)$ is the cumulative distribution function of a Beta random variable with parameter (a,b), then $Y = F^{-1}(\Phi(\epsilon); a, b)$ is a Beta random variable with parameter (a,b). Similarly to Guolo and Varin (2014), we consider a discrete time stochastic process $\boldsymbol{\epsilon} = (\epsilon_t)_{t\geq 1}$ defined as

$$\epsilon_1 \sim \mathrm{N}(0,1) \quad \text{and} \quad \epsilon_t = \psi \epsilon_{t-1} + \eta_t \qquad t \geq 2 \tag{3}$$

where $\psi \in (-1,1)$ and $(\eta_t)_{t>1}$ are independent and identically distributed (i.i.d.) as $\mathrm{N}(0, 1-\psi^2)$. That is, $\boldsymbol{\epsilon}$ is an autoregressive stochastic process of order 1 with parameter ψ and $\epsilon_t \sim \mathrm{N}(0,1)$; for brevity, $\boldsymbol{\epsilon} \sim \mathrm{AR}(1;\psi)$. Let $(\boldsymbol{\epsilon}_l)_{l\geq 1}$ be i.i.d. such that $\boldsymbol{\epsilon}_l \sim \mathrm{AR}(1;\psi)$ and let define

$$\xi_{tl} = F^{-1}(\Phi(\epsilon_{tl}); a, b), \tag{4}$$

for any $t \geq 1$ and $l \geq 1$. Because of the autoregressive structure of $\boldsymbol{\epsilon}_l$, ξ_{tl} depends on $\xi_{(t-1)l}$, with the parameter ψ controlling the dependence among the ξ_{tl}'s. In particular, for any fixed $l \geq 1$, the assumption $\psi = 0$ corresponds to the assumption of independence among the ξ_{tl}'s. Furthermore, for every $t \geq 1$, ξ_{tl} is independent of ξ_{th} if $l \neq h$, and they have the same distribution.

Let G denote a DP on $\Theta \subset \mathbb{R}^p$ with parameter (M, G_0), where G_0 is a non-atomic (base) distribution on Θ and $M > 0$ is the scale parameter; for brevity $G \sim \mathrm{DP}(M, G_0)$. It is known from the work of Sethuraman (1994) that $G = \sum_{h\geq 1} w_h \delta_{\theta_h}$ where: i) $(w_h)_{h\geq 1}$ are such that $w_1 = \xi_1$ and $w_h = \xi_h \prod_{1\leq l\leq h-1}(1-\xi_l)$ for $h > 1$, with $(\xi_l)_{l\geq 1}$ being i.i.d. as a Beta distribution with parameter $(1, M)$; ii) $(\theta_h)_{h\geq 1}$ are i.i.d. as G_0, and independent of $(\xi_l)_{l\geq 1}$. This is known as stick-breaking construction of $G \sim \mathrm{DP}(M, G_0)$. We exploit (4) to generalize the stick-breaking construction of the DP in order to define a sequence of dependent random probability measures $(G_t)_{t\geq 1}$, with t being a discrete time, such that $G_t \sim \mathrm{DP}(M, G_0)$. As such, let

$$G_t = \sum_{h\geq 1} w_{th} \delta_{\theta_h}, \tag{5}$$

where $w_{t1} = \xi_{t1}$ and $w_{th} = \xi_{th}\prod_{l=1}^{h-1}(1-\xi_{tl})$, for $h > 2$, with $(\xi_{tl})_{t\geq 1, l\geq 1}$ and $(\theta_h)_{h\geq 1}$ such that: i) $(\xi_{tl})_{t\geq 1, l\geq 1}$ are distributed as in (4), with parameters $a = 1$ and $b = M$; ii) $(\theta_h)_{h\geq 1}$ are i.i.d. with common distribution G_0, and independent of $(\xi_{tl})_{t\geq 1, l\geq 1}$. The dependent random probability measure $(G_t)_{t\geq 1}$ is referred to as the autoregressive DP of order 1; for brevity, $(G_t)_{t\geq 1} \sim$ AR1-DP(ψ, M, G_0). $(G_t)_{t\geq 1}$ is a dependent DP in the sense of MacEachern (2000). It is also an example of the single atom dependent DPs of Barrientos, Jara and Quintana (2012).

Since $\epsilon_{tl} \sim \mathrm{N}(0, 1)$ and F is the cumulative distribution function of a Beta random variable with parameter $(1, M)$, then ξ_{tl} is a Beta random variable with parameter $(1, M)$. This implies that if $(G_t)_{t\geq 1} \sim$ AR1-DP(ψ, M, G_0) then $G_t \sim$ DP(M, G_0) for any $t \geq 1$. Observe that the ξ_{tl}'s in (4), whose dynamics in $t \geq 1$ is driven by (3), induces a dynamics in the sequence of random probability measures $(G_t)_{t\geq 1}$. Most importantly, for every $l \geq 1$ the stochastic process $(\xi_{tl})_{t\geq 1}$ inherits the same autoregressive (order 1) Markov structure of each stochastic process $\boldsymbol{\epsilon}_l$. Therefore $(G_t)_{t\geq 1} \sim$ AR1-DP(ψ, M, G_0) has an autoregressive (order 1) Markov structure, with the parameter ψ controlling the dependence among the G_t's. In particular, the assumption $\psi = 0$ corresponds to independence among the G_t's. A natural extension of $(G_t)_{t\geq 1} \sim$ AR1-DP(ψ, M, G_0) arises by setting F to be the cumulative distribution function of a Beta random variable with parameter (a_t, b_t), for any $t \geq 1$. If we further assume that $a_t = a$ and $b_t = b$, for any $t \geq 1$, then $(G_t)_{t\geq 1}$ is such that G_t is the generalized DP of Hjort (2000). Although these generalization may be of interest to obtain more flexible prior distributions for dependent random probability measures in discrete time, here we focus on the AR1-DP prior.

For ease of explanation, hereafter we drop the sub-index l. From $\xi_1 = 1-(1-\Phi(\epsilon_1))^{1/M}$ (see Equation (4)) we write $\epsilon_1 = \Phi^{-1}(1-(1-\xi_1)^M)$, and from $\xi_2 = 1-(1-\Phi(\epsilon_2))^{1/M}$, using (3) and the expression of ϵ_1, we have $\xi_2 = 1-[1-\Phi(\psi\Phi^{-1}(1-(1-\xi_1)^M)+\eta_2)]^{1/M}$, where $\eta_2 \sim \mathrm{N}(0, 1-\psi^2)$. Accordingly, the conditional distribution of ξ_2 given ξ_1 coincides with the distribution of $1-(1-\Phi(Z))^{1/M}$, where $Z \sim \mathrm{N}(\psi\Phi^{-1}(1-(1-\xi_1)^M), 1-\psi^2)$. Along similar lines, the conditional distribution of ξ_t given ξ_{t-1} coincides with the distribution of

$$1-(1-\Phi(Z))^{1/M}, \tag{6}$$

where $Z \sim \mathrm{N}\left(\psi\Phi^{-1}\left(1-(1-\xi_{t-1})^M\right), 1-\psi^2\right)$. Equation (6) is crucial for sampling from $(G_t)_{t\geq 1} \sim$ AR1-DP(ψ, M, G_0). Figure 1 displays the conditional density function of ξ_2 given $\xi_1 = 0.5$ (left) and given $\xi_1 = 0.9$ (right) for different values of ψ and $M = 1$. Our construction is flexible, allowing for different shapes of the distribution. In particular, for $\psi = 0$ the conditional distribution coincides with the marginal distribution and it is a Uniform distribution.

2.2. *Posterior analysis.* To better describe the particle MCMC algorithm for posterior inference under the AR1-DP mixture model, it is useful to recall the probabilistic sampling structure of the DP. Because of the discreteness of the DP (Blackwell and MacQueen, 1973), a random sample $(\theta_1, \ldots, \theta_n)$ from $G \sim$ DP(M, G_0) features $1 \leq K_n \leq n$ distinct types, labelled by $(\theta_1^*, \ldots, \theta_{K_n}^*)$, with corresponding frequencies $(N_{1,n}, \ldots, N_{K_n,n})$ such that $1 \leq N_{i,n} \leq n$ and $\sum_{1\leq i\leq K_n} N_{i,n} = n$. In other terms, the random sample $(\theta_1, \ldots, \theta_n)$ induces a random partition Π_n of the set $\{1, \ldots, n\}$ into K_n blocks with sizes $(N_{1,n}, \ldots, N_{K_n,n})$. In particular, the probability of any partition of $\{1, \ldots, n\}$ having k blocks with frequency counts $(n_1, \ldots, n_k)$ is

$$p_{k,n}(n_1, \ldots, n_k) = \frac{M^k}{\prod_{i=0}^{n-1}(M+i)}\prod_{i=1}^{k}(n_i-1)!. \tag{7}$$

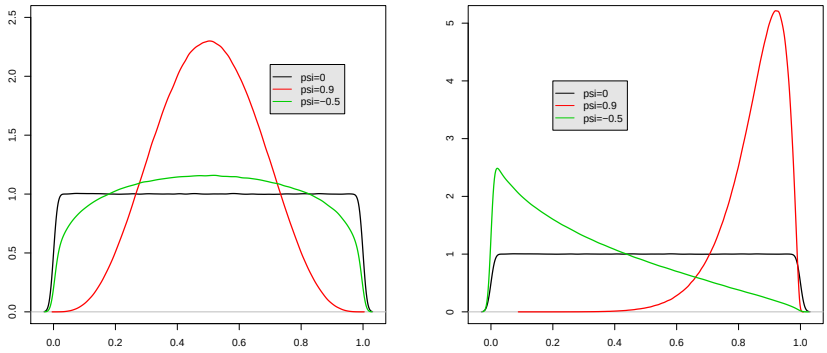

Fig 1: Conditional density of ξ_2 given $\xi_1 = 0.5$ (left) and given $\xi_1 = 0.9$ (right) for some values of ψ and $M = 1$.

See Antoniak (1974) for a detailed account of (7). In our context, a random sample $(\theta_{t1}, \ldots, \theta_{tn})$ from G_t, for $t = 1, \ldots, T$ and with $(G_t)_{t\geq 1} \sim$ AR1-DP(ψ, M, G_0), induces a collection $\{\Pi_{1,n}, \ldots, \Pi_{T,n}\}$ of dependent random partitions of $\{1, \ldots, n\}$. Due to the definition of AR1-DP(ψ, M, G_0), $\Pi_{t,n}$ is distributed as (7), for any $t = 1, \ldots, T$. The number of blocks/clusters of $\Pi_{t,n}$, denoted by $K_{t,n}$, at each time t is a random variable that is stationary, so that its prior marginal distribution will not change with t. We omit the subscript t and the subscript n when there is no chance of misunderstanding. Recall that the prior mean of the number K of clusters for any t, given the total mass M, is $\sum_{1\leq i\leq n} M/(M + i - 1)$ (Antoniak, 1974).

Now, we consider the problem of sampling from the posterior distribution of the AR1-DP mixture model. The design of a Gibbs sampler for such a problem is straightforward, once we truncate the infinite series (5) to J terms and we introduce allocation variables s_{tj} for each of the latent variable θ_{tj}. In particular, by using a latent variable representation, we can write

$$
\begin{aligned}
Y_{tj} \mid s_{tj}, \theta_{tj} &\overset{\text{ind}}{\sim} k(\cdot\,; \theta_{s_{tj}}) \quad j = 1, \ldots, n \\
s_{tj} \mid \mathbf{w}_t &\overset{\text{iid}}{\sim} \sum_{h=1}^{J} w_{th}\delta_h \quad j = 1, \ldots, n \qquad (8) \\
\theta_{tj} &\overset{\text{iid}}{\sim} G_0,
\end{aligned}
$$

where, for any $t = 1, 2, \ldots, T$,

$$
w_{t1} = \xi_{t1} \quad \text{and} \quad w_{tj} = \xi_{tj} \prod_{h=1}^{j-1}(1 - \xi_{tl}) \quad j = 2, \ldots, J-1, \qquad (9)
$$

with $w_{tJ} = 1 - \sum_{h=1}^{J-1} w_{th}$ and

$$
\xi_{tj} = 1 - (1 - \Phi(\epsilon_{tj}))^{1/M} \quad j = 1, \ldots, J-1 \qquad (10)
$$

$$
\epsilon_{1j} \overset{\text{iid}}{\sim} \text{N}(0, 1) \quad j = 1, \ldots, J-1 \qquad (11)
$$

$$
\epsilon_{tj} \mid \epsilon_{t-1,j}, \psi \overset{\text{ind}}{\sim} \text{N}(\psi\epsilon_{t-1,j}, 1 - \psi^2) \quad j = 1, \ldots, J-1. \qquad (12)
$$

Here $\mathbf{s}_t = (s_{t1}, \ldots, s_{tn})$ is the allocation vector at time t, whose elements denote the mixture component to which the elements of the sample $(Y_{t1}, \ldots, Y_{tn})$ are allocated at time

t; $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{tJ-1})$ is the latent autoregressive process; $\boldsymbol{\theta}_t = (\theta_{t1}, \ldots, \theta_{tJ})$ are the component-specific parameters and $\boldsymbol{w}_t = (w_{1t}, \ldots, w_{tJ-1})$ are the weights of the components in G_t (see Section 2.1). Therefore, the unknown parameters of the AR1-DP mixture model are $\{\boldsymbol{\theta}_t, \mathbf{s}_t, \mathbf{w}_t, \boldsymbol{\epsilon}_t, \psi, M\}$. The BNP mixture model defined in (8)-(12) not only provides a flexible tool for density estimation of discrete-time data, but also for time dependent cluster estimation. Furthermore, the proposed modelling strategy can be employed to model multiple time series, as an alternative to popular approaches which specify a distribution for the vector of observations of each subject. In this setting it is difficult to achieve dynamic clustering, with cluster membership changing over time. For a different proposal for time dependent clustering see Page, Quintana and Dahl (2021), who explicitly model dependence in a sequence of partitions by introducing auxiliary variables which identify which subject at time $t-1$ is considered for possible cluster reallocation at time t.

.

We outline the MCMC scheme for sampling from the posterior distributions of $\{\boldsymbol{\theta}_t, \mathbf{s}_t, \mathbf{w}_t, \boldsymbol{\epsilon}_t, \psi, M\}$. Further details of the algorithm can be found in Section 1 of the Supplementary Material. Note that while $\boldsymbol{\theta}$ is the same for each time period, the vector $\mathbf{s}_t$ changes over time so that individuals can change clusters. Formally, the main steps of the MCMC algorithm are the following:

1. *sampling $\boldsymbol{\theta}$ given the rest*: this step requires to sample the values of θ_l's corresponding to non-allocated (empty) component from the base distribution G_0, i.e. $\theta_l \overset{\text{iid}}{\sim} G_0$, while update θ_l corresponding to the allocated components from the following conditional distribution
$$p(\theta_l \,|\, \text{rest}) \propto G_0(\mathrm{d}\theta) \prod_{(t,j):s_{tj}=l} f(y_{tj}\,;\, \theta_{tj})$$
2. *sampling $(\mathbf{s}_1, \ldots, \mathbf{s}_T)$ given the rest*: the distribution of $(\mathbf{s}_1, \ldots, \mathbf{s}_T)$ can be factorized into the product of the distributions of each $\boldsymbol{s}_t$ given the rest; this is because, given $(\mathbf{w}_1, \ldots, \mathbf{w}_T)$, the allocations at different times are conditionally independent; in particular, for each t, j we have
$$p(s_{tj} \,|\, \text{rest}) \propto w_{tl} f(y_{tj}\,;\, \theta_l)$$
3. *sampling $\{\psi, (\mathbf{w}_1, \ldots, \mathbf{w}_T)\}$ given the rest*: this step requires to sample from the conditional distribution of $\{\psi, (\mathbf{w}_1, \ldots, \mathbf{w}_T)\}$ given $(\mathbf{s}_1 \ldots, \mathbf{s}_T)$, and we make use of a particle MCMC update, which is discussed more in details in Section 1 of the Supplementary Material.

The AR1-DP mixture model allows to accommodate a variety of temporal dynamic behaviours, thus defining a flexible class of time-evolving random density functions. Figure 2 displays some realizations of the AR1-DP(ψ, M, G_0) mixture model with the specification of a Gaussian kernel $k(\cdot;\cdot)$, for different values of ψ, and when: i) $T = 4$ and $M = 2$; ii) G_0 is a Uniform distribution with parameter $(-20, 20)$; iii) the truncation level in (8) is fixed at $J = 8$. See also Figure 13 in Section 3 of the Supplementary Material for the distribution of the Hellinger distances between $f_t, t = 2, 3, 4$ and f_1, defined as in (2). In particular, the Hellinger distance shows a time-dependent behaviour which is what we expect from an AR1 process.

2.3. *Competitor models.* Taddy (2010) introduces a dependent DP prior for modelling (discrete) time series of marked spatial point patterns. The likelihood function is assumed to factorize in two independent components: i) the integrated intensities of the Poisson processes, which are modelled as dynamic linear models; ii) a collection of the density functions

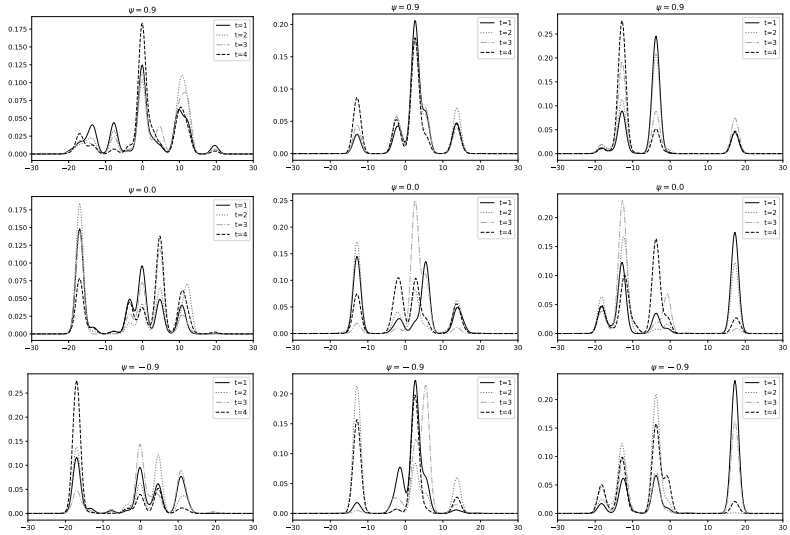

Fig 2: Realizations of f_t, for $t = 1, 2, 3, 4$ and for some values of ψ.

$(f_t)_{t\geq 1}$, which are modelled via a Bayesian nonparametric mixture model with a dependent DP prior. Specifically, the dependent DP prior of Taddy (2010) is defined as the distribution of $(G_t)_{t\geq 1}$ such that

$$G_t = \sum_{h\geq 1} \xi_{th} \prod_{l=1}^{h-1} (1 - \xi_{tl}) \delta_{\theta_h},$$

where

$$\xi_{tl} = 1 - u_{tl}(1 - w_{tl}\xi_{(t-1)l}) \tag{13}$$

with: i) u_{tl} and w_t being Beta random variables with parameter $(M, 1-\psi)$ and $(\psi, 1-\psi)$, respectively, for any $t \geq 1$, with $M > 0$ and $0 < \psi < 1$; ii) $(\theta_h)_{h\geq 1}$ being random variables i.i.d. with common distribution G_0, and independent of $(\xi_{tl})_{t\geq 1, l\geq 1}$. Accordingly, ξ_{tl} is distributed as a Beta distribution with parameter $(1, M)$ and hence $G_t \sim \text{DP}(M, G_0)$. The dependent DP prior of Taddy (2010) introduces only the additional prior parameter $0 < \psi < 1$, which accounts for modeling dependence over time. The correlation between ξ_{tl} and $\xi_{(t-k)l}$ is $(\psi M/(1 + M - \psi))^k > 0$, which rules out negative correlation among the random probability measures G_t's.

DeYoreo and Kottas (2018) develop a dependent DP prior for temporal dynamic ordinal regressions. The focus is on modelling the time relationships between the maturity of fishes (the ordinal response) and the age and length of fishes. In particular, the density functions for the maturation, length and age are modelled via a Bayesian nonparametric dynamic mixture model with a dependent DP prior. Differently from our AR1-DP prior, the prior of DeYoreo and Kottas (2018) has both random atoms and random weights depending on $t \geq 1$. Specifically, the dependent DP prior of DeYoreo and Kottas (2018) is defined as the distribution of $(G_t)_{t\geq 1}$ with

$$G_t = \sum_{h\geq 1} \xi_{th} \prod_{1\leq l\leq h-1} (1 - \xi_{tl}) \delta_{\theta_{t_h}},$$

where

$$\xi_{tl} = 1 - \exp\left\{-\frac{\zeta_l^2 + \eta_{tl}^2}{2M}\right\}, \tag{14}$$

and $(\eta_{tl})_{t\geq 1} \overset{\text{iid}}{\sim} \text{AR}(1,\psi)$, with $\psi \in (-1,1)$, $M>0$ and $\zeta_l \overset{\text{iid}}{\sim} N(0,1)$. This construction implies that $1-\exp\{-\frac{\zeta_l^2+\eta_{tl}^2}{2M}\}$ is distributed according to a Beta random variable with parameter $(1,M)$, and hence $G_t \sim \text{DP}(M,G_0)$. In particular, DeYoreo and Kottas (2018) show that, since η_{tl} enters squared in the stick-breaking weights ξ_{tl}'s, the correlation between ξ_{t-kl} and ξ_{tl} depends on the factor $(\psi^2)^k$ and the correlation between G_t and G_{t+1} is always positive. Furthermore, $M \geq 1$ implies that 0.5 is a lower bound on the correlation between ξ_{tl} and ξ_{t-kl}, for any $\psi \in (-1,1)$, and that such a peculiar issue may be overcome by time-varying locations.

We remark that the works of Taddy (2010) and DeYoreo and Kottas (2018) focus on dynamic density estimation, and they do not consider the problem of dynamic clustering. In particular, Taddy (2010) states that the dynamic clustering produced by his model is not robust, and very different clustering may corresponds to relatively similar predictive distributions. However, the nonparametric priors developed in Taddy (2010) and DeYoreo and Kottas (2018), as well as our prior, are stationary, so that, once sample from G_t, features like the number of unique values in the sample, will have the same marginal distribution under the different priors.

To highlight its flexibility of the AR1-DP mixture model, we make use of a simulation study to compare it with Taddy (2010) and DeYoreo and Kottas (2018). We generate data for $T=10$ time periods from mixtures of Gaussian distributions indexed by time and varying cluster assignments over time.

Specifically, at time $t=1$ we assume that the first 95 individuals are assigned to cluster 1 and sample $Y_{tj} \overset{\text{iid}}{\sim} N(\tilde{\mu}_1, \tilde{\sigma}_1^2)$ for $j=1,\ldots,95$, while 5 observations are assigned to cluster 2, and sample iid values from $N(\tilde{\nu}_1, \tilde{\tau}_1^2)$. At time $t=2,\ldots,10$, we assume the cluster allocation $s_{tj}=s_{t-1j}$ with probability 0.2 and $s_{tj}\neq s_{t-1j}$ with probability 0.8 (cluster allocations may assume only values in $\{1,2\}$); hence, at each $t\geq 2$ the number of individuals who change the cluster allocation from previous time is a random variable $\tilde{K}_t$ with binomial distribution with parameters $n=100$ and success probability 0.8. Then, for each $t\geq 2$ we simulate iid values from $N(\tilde{\mu}_t, \tilde{\sigma}_t^2)$ for individuals such that $s_{tj}=1$ and iid values from $N(\tilde{\nu}_t, \tilde{\tau}_t^2)$ for individuals such that $s_{tj}=2$. Values of means $\tilde{\mu}_t$, $\tilde{\nu}_t$ and standard deviations $\tilde{\sigma}_t$ and $\tilde{\tau}_t$ are reported in Table 1. We fit the AR1-DP mixture model to the simulated data:

$$
\begin{aligned}
Y_{tj} \mid (\mu_{tj},\tau_{tj}) &\overset{\text{ind}}{\sim} \text{N}(\mu_{tj}, (\lambda\tau_{tj})^{-1}) \quad j=1,\ldots,100 \\
(\mu_{tj},\tau_{tj}) \mid G_t &\overset{\text{iid}}{\sim} G_t \quad t=1,2,3,\ldots,10 \\
(G_t)_{t\geq 1} &\sim \text{AR1-DP}(\psi, M_t, G_0) \\
M_t | b &\overset{\text{iid}}{\sim} \text{Gamma}(3,b) \quad b \sim \text{Gamma}(300,50)
\end{aligned}
\tag{15}
$$

where G_0 is a Gaussian-Gamma distribution with parameter $(\mu_0,\lambda,\alpha,\beta)$, i.e. $\mu|\tau \sim \text{N}(\mu_0, \frac{1}{\lambda\tau})$ and $\tau \sim \text{Gamma}(\alpha,\beta)$ with $\mathbb{E}(\tau)=\alpha/\beta$. We set $\mu_0=\bar{Y}$, $\lambda=0.01$, $\alpha=5$, $\beta=1$, $J=20$ and $R=500$. The (prior) expected number K of clusters at time t is 3. We run the MCMC algorithm for 40,000 iterations, discarding the first 20,000 iterations as burn-in and thinning every 20. Therefore we obtain a total of 1,000 samples. The estimated number of clusters over time, obtained by minimizing the posterior expectation of variation of information (VI) loss function (Wade and Ghahramani, 2018), as implemented in the R package `BNPmix`, is 1, 1, 1, 2, 1, 1, 1, 1, 2, 1 respectively. The posterior distribution of ψ is displayed in Figure 3, with mean of ψ equal to 0.0624, and posterior probability of $\psi \leq 0$ equal to 0.376. Posterior co-clustering probabilities at $t=1$, i.e. the probability that two items in the sample are assigned to the same cluster, are displayed in Figure 4 (left panel). Estimated density functions overlapped with the true density functions are displayed in Figure 5.

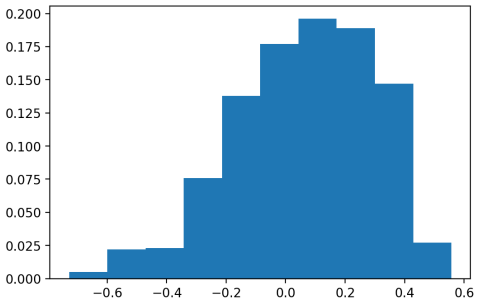

Fig 3: Marginal posterior distribution of ψ.

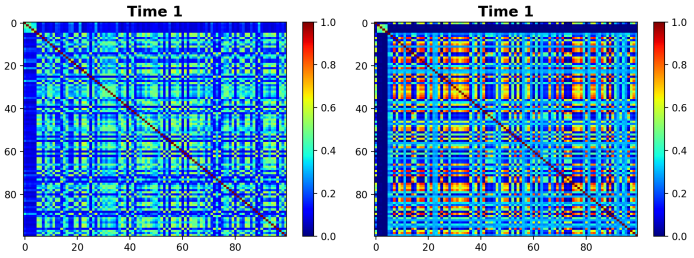

Fig 4: Simulated Example: Posterior co-clustering probability plots at $t = 1$ under the AR1-DP model (left) and Taddy (2010) (right).

TABLE 1

Simulated data: Mean and standard deviations (SD) for mixtures of Gaussian distributions indexed by time periods $t = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$

	$N(\mu_t, \sigma_t)$		$N(\nu_t, \tau_t)$	
t	μ_t	σ_t	ν_t	τ_t
1	-3	1.5	3	1.5
2	-6	2.5	2	2.5
3	-4	2	4	2
4	-2	1	6	1
5	-3	3	3	3
6	-3	1.5	3	1.5
7	-6	2.5	2	2.5
8	-4	2	4	2
9	-2	1	6	1
10	-3	3	3	3

We fit the models of Taddy (2010) and DeYoreo and Kottas (2018) to the same simulated data. For both the priors proposed by Taddy (2010) and DeYoreo and Kottas (2018), we assume that G_0 is a Gaussian-Gamma distribution with parameter $(\mu_0, \lambda, \alpha, \beta)$, i.e. $\mu|\tau \sim \text{N}(\mu_0, \frac{1}{\lambda\tau})$ and $\tau \sim \text{Gamma}(\alpha, \beta)$ with $\mathbb{E}(\tau) = \alpha/\beta$. We set $\mu_0 = \bar{Y}$, $\lambda = 0.01$, $\alpha = 5$, $\beta = 1$. For the prior of Taddy (2010) we assume $M \sim \text{Gamma}(3, 6)$ and $\psi \sim \text{Uniform}(0, 1)$, whereas for the prior of DeYoreo and Kottas (2018) we assume $\psi \sim \text{Uniform}(0, 1)$, $M \sim \text{Gamma}(3, 6)$ and $\tau_m \overset{\text{iid}}{\sim} \text{Gamma}(5, 1)$. In this case, we also assume time-varying location, i.e. we fit the following model: for all t, observations Y_{tj}, for $j = 1, \ldots, n$, are i.i.d. according

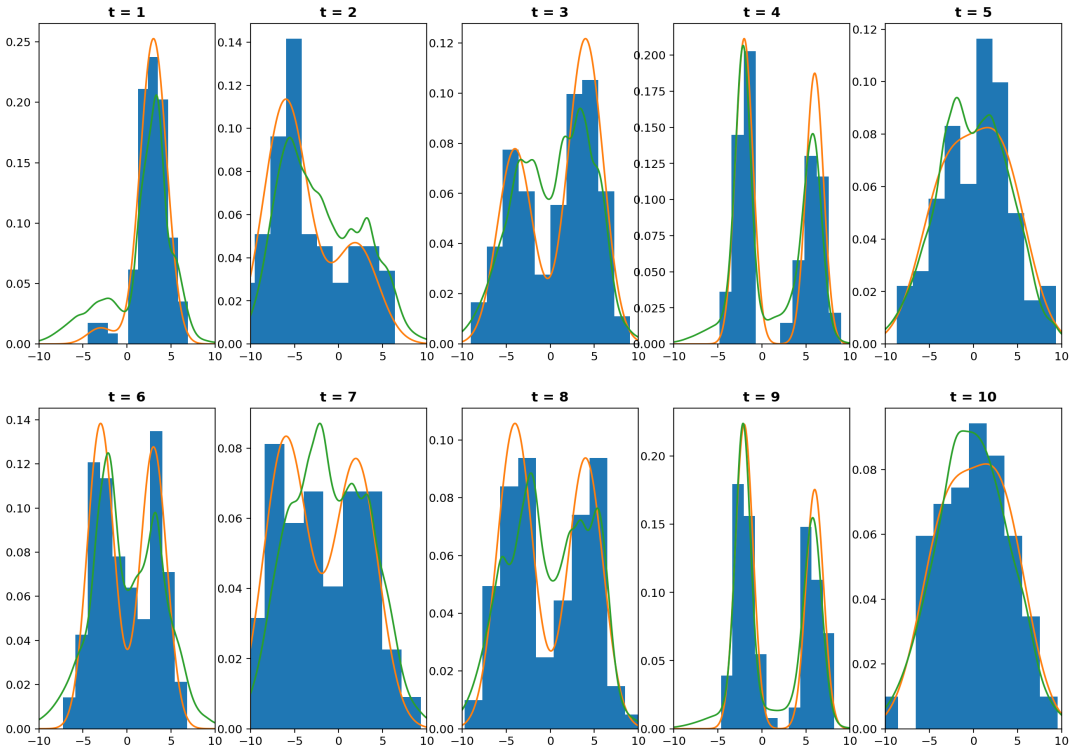

Fig 5: Posterior predictive densities (green lines) versus true densities (orange lines) with data (blue bar). *Colors are visible in the online version.*

to $f(y_{tj} \mid G_t)$, where

$$f(y_{tj}|G_t) = \sum_{m=1}^{\infty} p_{m,t} \mathrm{N}(y_{tj}|\mu_{m,t}, \tau_m^{-1}) \tag{16}$$

with

$$\mu_{m,0} \sim \mathrm{N}(0, 10), \quad \mu_{m,t}|\mu_{m,t-1} \sim \mathrm{N}(\theta\mu_{m,t-1}, 10)), \quad \tau_m \overset{\text{iid}}{\sim} \text{Gamma}(2, 2) \tag{17}$$

and the positive parameter θ being random. We assume a priori independence, when not differently specified, among blocks of parameters. Posterior computations are performed by truncating the infinite stick-breaking representation, as for our model, at a truncation level $J = 20$. Observe that the model of DeYoreo and Kottas (2018) assumes a positive random ψ, and that the parameter θ represents the autocorrelation parameter between the cluster locations.

Both prior specifications for the parameter θ discussed by DeYoreo and Kottas (2018) are considered, i.e. $\theta \sim \text{Uniform}(0, 1)$ and $\theta \sim \text{Uniform}(-1, 1)$, respectively. In both cases, the number of estimated parameters is only one, so that the corresponding co-clustering plots do not provide any information. We noted that, to obtain more than one estimated cluster at each time point, we need to force the parameter τ in the baseline measure G_0 to be highly concentrated on a large value, showing that the model proposed in DeYoreo and Kottas (2018) as specified in (16)-(17) is less flexible than ours. The model in Taddy (2010) gives a larger (than ours) number of estimated clusters at each time t (7, 10, 8, 3, 10, 9, 12, 9, 2, 7), though the density estimates (not shown here) looks very similar to ours (and they follow the histograms). In general, models that give an overestimated number of clusters produce better

density estimates, and this seems to be the case for two times ($t = 3, 7$) for our model. In conclusion, this simulated example shows that the version of the model by DeYoreo and Kottas (2018) we consider here cannot capture the negative correlation of the dynamic location specific parameters. The prior process proposed by Taddy (2010) seems to overestimate the number of clusters. However, it must be noted that, if we fit the model of DeYoreo and Kottas (2018) using our MCMC scheme (instead of the one described in the original paper), this latter approach is more computationally efficient than our approach and the one of Taddy (2010). In scenarios 3 and 8 in Section 2 in Supplementary Material we include a comparison of cluster and density estimates between our model and the models of Taddy (2010) and DeYoreo and Kottas (2018).

3. Application to gender stereotypes. We apply the AR1-DP mixture model (1) to time-dependent data for adjective embedding bias and for occupation embedding bias. We study how gender stereotypes, with respect to adjectives and occupations, change over time in the 20th and 21th centuries in the United States. We make use of word embeddings (Garg et al., 2018) to measure the gender bias. We consider embeddings trained on Corpus of Historical American English (COHA) (Hamilton, Leskovec and Jurafsky, 2016) for eleven decades $t = 1900, 1910, 1920, \ldots, 2000$. These embeddings are applied to lists of words from Garg et al. (2018), representing each gender (men and women) and neutral words (occupations and adjectives). Since there were ties in the original bias data, we jittered the data by adding a zero-mean Gaussian noise before standardization. This leads to data for (standardized) adjective and occupation embedding biases for women, for each word in the corresponding list. For each time t, we obtain two (unidimensional) datasets: i) the occupation biases $\{y_{tj}, j = 1, \ldots, n_O\}$, with $n_O = 76$; ii) the adjective biases $\{z_{tl}, l = 1, \ldots, n_A\}$, with $n_A = 230$). See Appendix A for details. A negative value of the bias means that the embedding more closely associates the word with men, because of the distance closer to men than women. Gender bias corresponds to either negative or positive values of the embedding bias. Hereafter, we write that there is a "bias against women" when the value of the embedding bias is negative.

We model occupation embedding biases $\{y_{tj}, j = 1, \ldots, n_O\}$ and adjective embedding biases $\{z_{tl}, l = 1, \ldots, n_A\}$ separately with the AR1-DP mixture model (1) with a Gaussian kernel. Specifically,

$$
\begin{aligned}
Y_{tj} \mid (\mu_{tj}, \tau_{tj}) &\overset{\text{ind}}{\sim} \text{N}(\mu_{tj}, (\lambda\tau_{tj})^{-1}) \quad j = 1, \ldots, n_O \\
(\mu_{tj}, \tau_{tj}) \mid G_t &\overset{\text{iid}}{\sim} G_t \quad t = 1900, 1910, 1920, \ldots, 2000 \\
(G_t)_{t\geq 1} &\sim \text{AR1-DP}(\psi, M, G_0).
\end{aligned} \tag{18}
$$

where G_0 is a Gaussian-Gamma distribution with parameter $(\mu_0, \lambda, \alpha, \beta)$, namely $\mu|\tau \sim \text{N}(\mu_0, \frac{1}{\lambda\tau})$ and $\tau \sim \text{Gamma}(\alpha, \beta)$ with $\mathbb{E}(\tau) = \alpha/\beta$. We assume the same model as in (18) for the Z_{tl}'s. We set $\mu_0 = 0$, $\lambda = 0.01$, $\alpha = 2$, $\beta = 1$. For the occupation biases, we assume $M \sim \text{Gamma}(4, 4)$, while for the adjective biases $M \sim \text{Gamma}(3, 5)$. In both cases the prior distribution of number K of clusters at time t concentrates most of its mass on $\{2, 3, \ldots, 10\}$, as suggested by the exploratory data analysis presented in Appendix A; the prior expectation of K is 4 in both cases. We have also performed a sensitivity analysis with respect to the choice of the prior distribution for M, slightly increasing the prior expectation of K, and the posterior inference is robust to this choice (results not shown). Posterior inference is performed through the MCMC algorithm described in Section 2.2. We set the number of iterations equal to 20,000, with a burn-in period of 10,000 iterations and thinning every 10 iterations. The truncation level J of the prior is fixed equal to the sample size in both the analysis.

3.1. *Posterior inference: occupational embedding bias.* Figure 6 (right-panel) shows that the posterior distribution of the parameter ψ puts all its mass on the positive real line, with a mode around 0.8 and posterior mean equal to 0.71. Figure 7 reports, only for three decades, the posterior co-clustering probabilities, i.e. the probability that the two items in the sample are assigned to the same cluster. From Figure 7, ~~the predominant colors in 1900 and 1950 are blue (low posterior co-clustering probability) and green and yellow (medium posterior co-clustering probability) but in different proportions, whereas the predominant colors in 2000 are yellow and green, indicating an overall posterior co-clustering probability larger than 1900 and 1950.~~ we have that the co-clustering probability are low (most values below 0.25) in 1900 and 1950, while they are higher in 2000 (most values being around 0.65). In particular, the decrease over time in the bias against women can be seen in Figure 8, where we plot posterior predictive densities of the occupational embedding bias for all the decades, though we comment only the predictive densities for $t = 1900, 1950, 2000$. This is in agreement with the data as it is evident from the boxplot of occupational bias embeddings over time in Figure 12 in Appendix A (the data in the figure are not standardized).

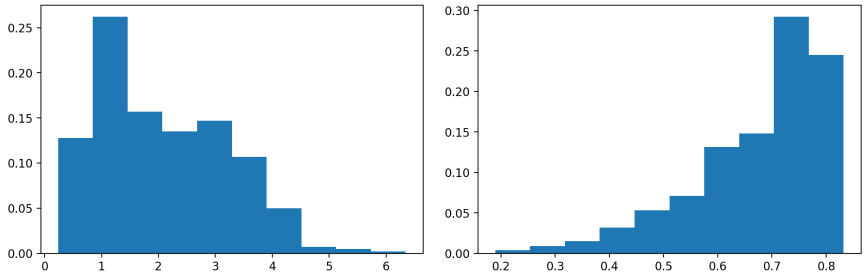

Fig 6: Marginal posterior distributions of M (left panel) and ψ (right panel) for occupational bias data.

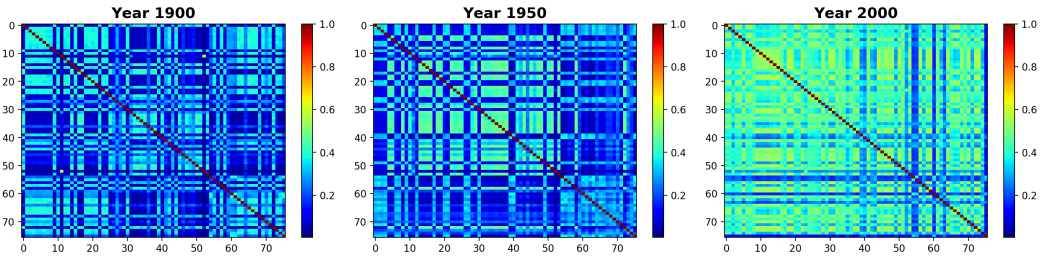

Fig 7: Co-clustering for occupational bias: $t = 1990$ (left panel), $t = 1950$ (center panel), $t = 2000$ (right panel).

From Figure 8, the estimated density function at time $t = 1900$ has two distinct peaks: a higher peak centered at a negative (gender) bias location, and a lower peak centered at a positive (gender) bias location. This means that in 1900, the fraction of man-biased jobs is larger than the fraction of woman-biased jobs. In 1950, the estimated density function still has two distinct peaks. While the locations of these two peaks remain about the same as in 1900, the fraction of man-biased jobs and woman-biased jobs changes over time. In particular, the fraction of man-biased jobs decreases with respect to the fraction of woman-biased jobs. That is, between 1900 and 1950 more occupations become gender neutral or woman-biased. However, for the year 2000, the estimated density function in Figure 8 has a single peak, on the positive real line, showing that most of jobs are neutral. Posterior predictive means increase over time: their value is -0.100, 0.120, 0.384 in 1900, 1950 and 2000, respectively.

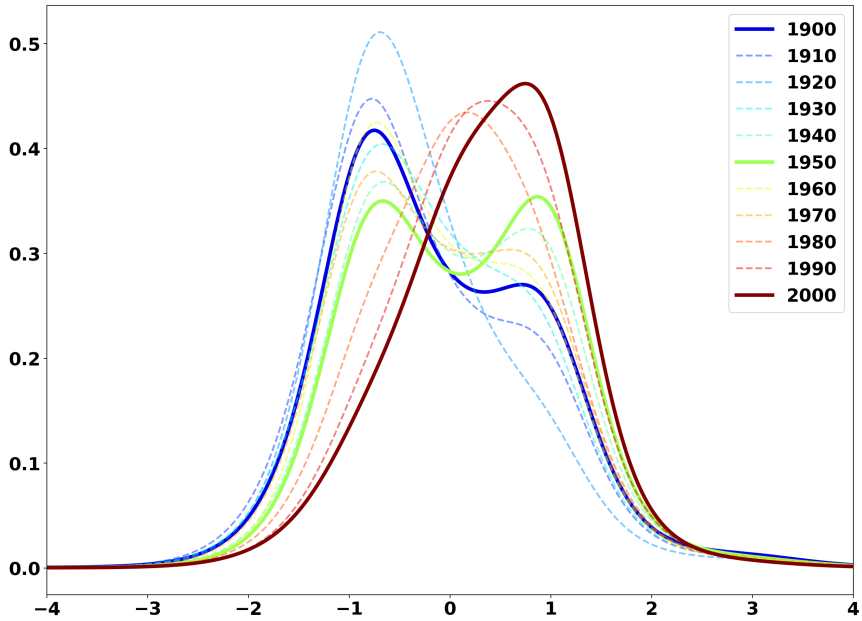

Fig 8: Posterior predictive densities of the occupational bias Y_t^{new}.

Table 2 shows estimated cluster configurations at times $t = 1900, 1910, 1920, \ldots, 2000$, discarding singletons. In particular, the table reports the clustering that minimizes the posterior expectation of variation of information (VI) loss function (Wade and Ghahramani, 2018), obtained via the R package `BNPmix`. Wade and Ghahramani (2018) show that cluster estimation using VI has the appealing property of penalising small clusters in comparison to Binder's loss (Binder, 1978), and therefore it leads to more interpretable estimated partition. The number of estimated clusters over the times is $2, 2, 4, 2, 4, 4, 2, 3, 1, 1, 1$, respectively. However, besides singletons in 1920, 1940, 1950 and 1970, the estimated number of clusters is always two except for the last three decades, in agreement with the estimated densities displayed in Figure 8. In order to better explain our estimates, clusters are interpreted by assigning them a "label" as follows *(see Table 2):* i) "man-cluster" ~~(green cells in Table 2)~~ (i.e. occupations in the cluster are biased against women) if the empirical mean of all the data points in the cluster is negative and zero is not within one (empirical) standard deviation from the mean; ii) "woman-cluster" ~~(blue cells)~~ (i.e. the occupations are biased in favour of women, or against men) if the empirical mean of all the data points in the cluster is positive and zero is not within one standard deviation from the mean itself. A cluster is "neutral" ~~(pink cells)~~ if zero is within one standard deviation from the empirical mean of all the data in the cluster. Mean, standard deviation, min, max and count on top of the cells in Table 2 refer to data without standardization.

Table 2 about here

As representative examples, we briefly discuss how few words change cluster over time, i.e. we discuss cluster dynamics. In particular, the occupation word "nurse", whose non-standardized embedding is always positive, stands as a singletons in 1920, 1940, 1950 and 1970, while it is always in a "neutral" cluster in 1900, 1910, 1930 and 1960. We also observe that in 1900, the occupation word "athlete" is associated with men, while in 1950 and 2000, it belongs to the "neutral-cluster". Of course, women athletes were very few in 1900, but their number started to increase during the 20th century. For instance, the number of Olympic women athletes increased from 65 at the 1920 Summer Olympics to 331 at the

1936 Summer Olympics. Data have been retrieved from the official website of the Olympic Games. See `https://www.olympic.org/` for details. For some words the dynamics of the corresponding cluster membership is less intuitive. For instance, the occupation word "midwife" belongs to " man " clusters in all the decades until 1970. This can be explained by the non-standardized data over time, since the bias embedding corresponding to "midwife" is always negative, but it increases sharply to a positive value only in the year 2000. On the other hand, the set of occupations that require a great amount of physical strength, such as "farmer" and "soldier", always belong to " man " clusters, and they co-cluster across all the decades.

We compare our results with those obtained from alternative approaches, i.e. model (16)-(17) by DeYoreo and Kottas (2018) as introduced in Section 2.3 (via our implementation) and model (10) in Page, Quintana and Dahl (2021) (as implemented in the R-package *drpm*). Cluster estimates depend on hyperparameter specification for all three models. We perform a sensitivity analysis to hyperparameter choice for the two competitors models. For instance, model (16)-(17) in DeYoreo and Kottas (2018) produces only one cluster for any time t, unless we *set* a prior which strongly favours a large number of clusters. Similar considerations apply to the model in Page, Quintana and Dahl (2021), using default (in the companion R package) hyperparameter values, which include a fixed value for the total mass parameter of the marginal random partition prior (unlike our model). Nevertheless, if we assume $E(\tau) = 5$ with variance 10^{-3} in (16)-(17) above (DeYoreo and Kottas, 2018), we find that the number of estimated cluster (under VI loss) is 7, 7, 5, 7, 8, 7, 6, 7, 7, 6, 6. Still, the difference between the associated cluster estimate and the one obtained with our model is evident, as quantified by the adjusted Rand index (Hubert and Arabie, 1985) among the two cluster estimates (0.54, 0.39, 0.11, 0.29, 0.36, 0.61, 0.41, 0.46, 0.00, 0.00, 0.00). Recall that the adjusted Rand index may assume also negative values, and it is bounded above by 1 which corresponds to the case of perfect agreement between two partitions (see the R package `mclust`, Scrucca et al., 2016). Under different hyperparameter settings (with total mass parameter of the marginal time-dependent random partition prior equal to 2) we find that the number of estimated clusters obtained with the approach of Page, Quintana and Dahl (2021) is 2 from 1900 to 1940, and then 1 from 1950 to 2000. Once again the associated cluster estimates, until 1960, are different from ours, as measured by the adjusted Rand index (0.14, -0.02, -0.03, 0.06, 0.05, 0, 0, 0, 1, 1, 1). Figure 11 displays the lagged ARI values for the cluster estimates for model (10) in Page, Quintana and Dahl (2021), DeYoreo and Kottas (2018) and our model. Our analysis highlights that the models by DeYoreo and Kottas (2018), and Page, Quintana and Dahl (2021) are more sensitive to the choice of hyperparameters than ours. in particular , the performance of model by Page, Quintana and Dahl (2021) is strongly affected by the choice of the mass parameter, which controls the number of clusters.

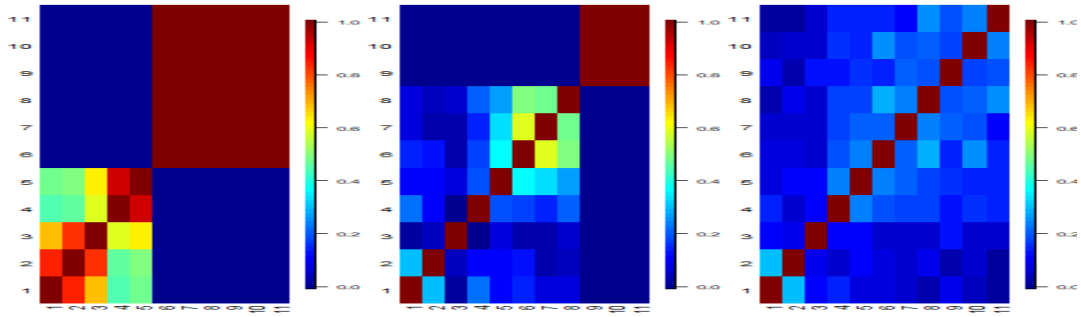

Fig 9: Lagged ARI values corresponding to Page, Quintana and Dahl (2021) (left) and our (center) and DeYoreo and Kottas (2018) (right) models. At each time point the partition was estimated using BNPmix R package based on VI loss.

3.2. *Posterior inference: adjective embedding bias.* We fit the unidimensional model (18), that is the unidimensional AR1-DP mixture model with a Gaussian kernel, to standardized adjective biases $\{z_{tl}, l = 1, \ldots, n_A\}$. From Figure 10, the estimated density function for the adjective embedding bias at time $t = 1900$ has three peaks: the highest peak is centered near the value -1.5, a small peak centered around the value -0.5 and a third peak centered at a positive value for the gender bias. This means that in 1900, the fraction of man-biased adjectives is larger than the fraction of neutral or woman-biased adjectives. In 1950 and 2000, the estimated density functions still have three peaks. While the location of these three peaks remain the same, the fractions of man-biased adjectives and woman-biased adjectives change. In particular, while the estimated density function in 1950 is very similar to that of $t = 1900$, the estimated density function in 2000 is different. Indeed in 2000, the peak centered near the value -1.5 becomes very small, while the peak centered around the value -0.5 increases notably and the peak on the positive location grows moderately. Posterior predictive means are 0.153, 0.165, 0.143 for $t = 1900, 1950, 2000$, respectively, showing a slow decrease over time.

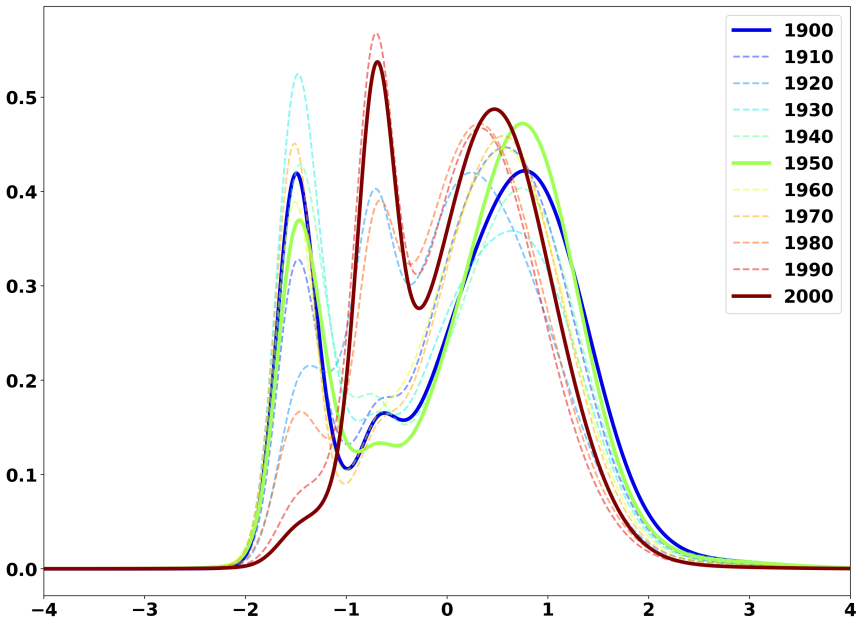

Fig 10: Posterior predictive densities of the adjective bias Z_t^{new}.

Table 3 shows the estimated cluster configurations at times $t = 1900, 1910, 1920, \ldots, 2000$, excluding all the singleton clusters. Also, the number of estimated clusters over the time periods is: 7, 8, 6, 9, 8, 7, 5, 5, 7, 5, 8. From Table 3, it is clear that bias tends to mitigate over the years. If we consider the total number of adjectives in " neutral " clusters, we go from 4 in 1900, to 141 in 1950, to 147 in 2000. However, the cluster estimate in 2000 does not succeed to identify a " woman " cluster, that is apparent from the density estimates in Figure 10. This may be explained looking at the boxplots for the last three decades (see Figure 13 in Appendix A), where data are still biased against women overall, but variability is smaller. For some of the adjective words, the inclusion in one group corresponds to conventional stereotypes. For example, the words "attractive", " charming " and " feminine ", often clustered together, are in " woman " clusters in almost all decades until 1970, and then they belong to " neutral " clusters. The words "immature", "enterprising" and "autocratic" are in " man " clusters for all the decades. Moreover, words such as "cool", "demanding", "relaxed" belong

to neutral clusters in most decades, e.g. in 1900, 1950 and 2000 as showed in Table 3. Furthermore, it is also clear from the table that some adjective words are clearly gender-stereotyped: "adaptable", "dependable", "inventive", "methodical", "resourceful" and "sociable" are always associated with men; analogously "rigid" is associated with women for most of the decades, being its embedding mostly positive. Figures 17 and 18, respectively, show the marginal posterior distribution of parameters ψ and M, as well as the posterior co-clustering probabilities.

Table 3 about here

We compare our clustering results with those obtained with the model of Page, Quintana and Dahl (2021). Similarly as for the application on occupation embedding bias, the analysis is very sensitive to hyperparameter choice. Using the same values as in the previous application, we get that the number of estimated clusters is on average smaller than in our case (specifically 2, 5, 5, 5, 5, 3, 2, 4, 5, 3, 2). This might be due to the fact that the *total mass* parameter M is random in our case, and fixed ($M = 2$) for the model of Page, Quintana and Dahl (2021). The difference between the two estimated partitions is substantial, with the adjusted Rand index at each time point equal to 0.43, 0.33 , 0.10, 0.51 , 0.46 , 0.00, 0.00, 0.59, 0.39, 0.00, -0.02 respectively. Figure 11 displays the lagged ARI values for the cluster estimates for model (10) in Page, Quintana and Dahl (2021) and our model.

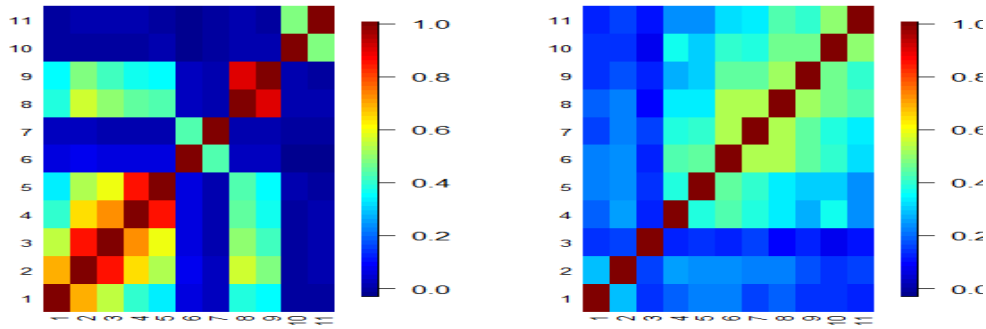

Fig 11: Lagged ARI values corresponding to Page, Quintana and Dahl (2021) (left) and our (right) models. At each time point the partition was estimated using BNPmix R package based on VI loss.

3.3. *Discussion on the dynamics of gender stereotypes.* We compare our posterior inference results with those obtained by Garg et al. (2018), as well as results found in Bolukbasi et al. (2016) and Caliskan, Bryson and Narayanan (2017) which are also discussed in Garg et al. (2018). Bolukbasi et al. (2016) propose a method for removing gender stereotypes in the embeddings, a problem we do not consider here, since word embeddings often show gender bias to a worrying extent. On the other hand, Caliskan, Bryson and Narayanan (2017) show that machine learning techniques as applied to ordinary human language results in human-like semantic biases. They replicate a range of known biases, as measured by the Implicit Association Test, arguing that language itself contains recoverable imprints of human historic biases. In particular, Caliskan, Bryson and Narayanan (2017) find female names are more associated with family than career words, as compared to male names, and that female words (such as "woman" and "girl") are more associated with the arts than with mathematics. From Garg et al. (2018) "gender bias, as seen through adjectives associated

with men and women, has decreased over time and that the women's movement in the 1960s and 1970s especially had a systemic and drastic effect in women's portrayals in literature and culture". The authors show that despite "women's occupation percentages are highly correlated with embedding gender bias", the embeddings generally reflect additional social stereotypes beyond what can be explained by occupation participation. In particular, Garg et al. (2018) prove "consistency" between occupational words' embeddings and women's occupation participation over two groups of decades, from 1900 to 1970 and from 1980 to 2000. The effects of women's movements, in particular on the language in the 1960-1970, is well documented (Rosen, 2013).

From the posterior predictive estimates in Figures 8 and 10 and from the cluster estimates in Tables 2 and 3, we see the same discontinuity over the time period under consideration. In particular, for the occupational words, Figure 8 shows that bias against women is high until 1970. Moreover, we observe two components in the density estimates for all decades including 1970, which then reduces to a single component after 1970. Bias against women decreases from 1970 and one "clear" component is enough to explain the predictive densities from 1980. This is also in agreement with the number of estimated clusters over time (beyond singletons) that is always 2 except in 1980, 1990 and 2000. Similar comments hold true for the adjective bias if we look at Figure 10, where the posterior predictive densities for all decades from 1900 to 1960 (included) are very similar, while from 1970, the peak centered near the value -1.5 becomes very small, almost undistinguishable, while the right peak seems to move toward zero. Remember that the parameter ψ controls the AR-1 dependence among the random mixing measures over time and hence only indirectly controls the dependence between data at time $t+1$ from data at time t. However, it is interesting to note that the posterior distributions of ψ for occupational bias (Figure 7, right panel), as well as for adjectives' bias (Figure 17, right panel), are in agreement with the empirical covariance between adjective bias reported in Table 4 of Garg et al. (2018) and between occupational bias (see Figure B.7 Garg et al. (2018)).

4. Conclusions. Motivated by recent methodological and computational developments on BNP analysis of time-dependent data, in this paper we investigated the use of BNP mixture modeling to study temporal dynamics of gender stereotypes. The AR1-DP provides a framework for both density estimation from discrete-time data and discrete-time dependent clustering of the subjects measured at different time points. Moreover, it can be employed for modeling multiple time series.

There has been a recent interest in the use of machine learning techniques, and in particular word embeddings in natural language processing, to quantify and compare gender and ethnic stereotypes over time. Here, we exploit word embeddings trained on COHA (Hamilton, Leskovec and Jurafsky, 2016) for eleven decades between 1900 to 2000, and a list of words provided by Garg et al. (2018), to obtain time-dependent data for (standardized) adjective and occupation embedding biases for women over the 20th and 21st centuries in the United States. Then, we developed a novel BNP dynamic mixture model, which is referred to as the AR1-DP mixture model, for modeling time-dependent adjective bias and for occupation bias data. The AR1-DP mixture model exploits both the discreteness of the DP prior and an autoregressive dependence structure among DP priors to provide a flexible and robust model for dynamic density estimation and for dynamic clustering of biases data in a hierarchical setting. Posterior inference is performed through a particle MCMC algorithm which is computationally efficient. The application of the AR1-DP mixture model to data of adjective and occupation embedding bias shows that our model is able to quantify historical trends of gender bias, and to identify how specific adjectives and occupations became more closely associated with certain populations over time. We find that the numbers of components (as shown in the posterior predictive densities) as well as the number of estimated

clusters changes between 1960 and 1970. Our analysis highlights a clear difference in behaviour between the last three decades and the previous ones, both in terms of density and cluster estimation. As such, our results complement existing literature on the topic.

Our work demonstrates that the AR1-DP mixture model is a powerful tool in the context of quantifying gender stereotypes through time-dependent data of adjective and occupation embedding biases. Besides this context, we believe that the flexibility of the AR1-DP prior may be usefully exploited to develop BNP dynamic mixture models in more general settings. In particular, our future research will focus on extensions of the AR1-DP mixture model in order to consider applications to time-dependent data with covariates and time-dependent data with spatial structure. Moreover, we will consider higher order dependence structures. It must be noted that the AR1-DP model assumes that the cluster memberships of subjects are conditionally i.i.d., given the mixing distribution G_t at time t. This implies, for example, that there is a positive probability that a word might jump from a man cluster at time t to a woman cluster at time $t+1$. As the analysis concerns the same set of words over time, we could impose some stronger dependency on cluster membership, at the cost of more expensive computations. Finally, it is also straightforward to extend our prior process to include dynamic locations as in DeYoreo and Kottas (2018) to allow for extra flexibility, if required by the application.

Acknowledgements. The authors are grateful to the Editor (Professor Brendan Murphy) and two anonymous Referees for their comments and corrections that allow to improve remarkably the paper. Stefano Favaro received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257. Stefano Favaro gratefully acknowledge the financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022.

REFERENCES

ALTMAN, M. (2003). Beyond trashiness: The sexual language of 1970s feminist fiction. *Journal of International Women's Studies* **4** 7–19.

ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** 269–342.

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2** 1152–1174.

ARBEL, J., MENGERSEN, K. and ROUSSEAU, J. (2016). Bayesian nonparametric dependent model for partially replicated data: the influence of fuel spills on species diversity. *The Annals of Applied Statistics* 1496–1516.

BARRIENTOS, A. F., JARA, A. and QUINTANA, F. A. (2012). On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Analysis* **7** 277–310.

BASOW, S. A. (1992). *Gender: stereotypes and roles*. Thomson Brooks/Cole Publishing Co, Belmont, CA).

BASSETTI, F., CASARIN, R. and LEISEN, F. (2014). Beta-product dependent Pitman–Yor processes for Bayesian inference. *Journal of Econometrics* **180** 49–72.

BINDER, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65** 31–38.

BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The annals of statistics* **1** 353–355.

BOLTANSKI, L. and CHIAPELLO, E. (2006). *The new spirit of capitalism*. Verso, London (UK).

BOLUKBASI, T., CHANG, K. W., ZOU, J. Y., SALIGRAMA, V. and KALAI, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems 29, Barcelona*.

CALISKAN, A., BRYSON, J. J. and NARAYANAN, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science* **356** 183–186.

CARON, F., DAVY, M. and DOUCET, A. (2007). Generalized Polya urn for time-varying Dirichlet process mixtures. In *Proc. 23rd Conf. on Uncertainty in Artificial Intelligence, Vancouver*.

CARON, F., DAVY, M., DOUCET, A., DUFLOS, E. and VANHEEGHE, P. (2008). Bayesian inference for linear dynamic models with Dirichlet process mixtures. *Signal Processing, IEEE Transactions on* **56** 71–84.

COATES, J. (2016). *Women, men and language: a sociolinguistic account of gender differences in language*. Routledge, London.

COLLOBERT, R., WESTON, J., BOTTOU, L., KARLEN, M., KAVUKCUOGLU, K. and KUKSA, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research* **12** 2493–2537.

DEYOREO, M. and KOTTAS, A. (2018). Modeling for dynamic ordinal regression relationships: an application to estimating maturity of rockfish in California. *Journal of the American Statistical Association* **113** 68–80.

DI LUCCA, M. A., GUGLIELMI, A., MÜLLER, P. and QUINTANA, F. A. (2013). A simple class of Bayesian nonparametric autoregression models. *Bayesian Analysis* **8** 63–88.

DUNSON, D. B. and PARK, J.-H. (2008). Kernel stick-breaking processes. *Biometrika* **95** 307–323.

DUNSON, D. B., PILLAI, N. and PARK, J.-H. (2007). Bayesian density regression Series B Statistical methodology.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1** 209–230.

GARG, N., SCHIEBINGER, L., JURAFSKY, D. and ZOU, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* **115** E3635–E3644.

GRIFFIN, J. E. and STEEL, M. J. (2006). Order-based dependent Dirichlet processes. *Journal of the American statistical Association* **101** 179–194.

GRIFFIN, J. E. and STEEL, M. F. (2011). Stick-breaking autoregressive processes. *Journal of econometrics* **162** 383–396.

GUOLO, A. and VARIN, C. (2014). Beta regression for time series analysis of bounded data, with application to Canada Google® Flu Trends. *The Annals of Applied Statistics* **8** 74–88.

GUTIÉRREZ, L., MENA, R. H. and RUGGIERO, M. (2016). A time dependent Bayesian nonparametric model for air quality analysis. *Computational Statistics & Data Analysis* **95** 161–175.

HAMILTON, W. L., LESKOVEC, J. and JURAFSKY, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.

HEWITT, N. A. (2012). Feminist frequencies: Regenerating the wave metaphor. *Feminist Studies* **38** 658–680.

HJORT, N. L. (2000). Bayesian analysis for a generalised Dirichlet process prior. *Technical Report, MatematiskInstitutt, Universitetet i Oslo*.

HOLMES, J. and MEYERHOFF, M. (2008). *The handbook of language and gender* **25**. John Wiley & Sons.

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *Journal of classification* **2** 193–218.

KASSAMBARA, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning* **1**. Sthda.

KAUFMAN, L. and ROUSSEEUW, P. J. (1990). Finding groups in data. *Hoboken: Wiley Online Library*.

LEWIS, M., COOPER BORKENHAGEN, M., CONVERSE, E., LUPYAN, G. and SEIDENBERG, M. S. (2020). What might books be teaching young children about gender? *Psychological science* 09567976211024643.

LO, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The annals of statistics* 351–357.

MACEACHERN, S. N. (2000). Dependent dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University* 1–40.

NICHOLSON, L. (2010). Feminism in 'Waves': Useful Metaphor or Not? *New Politics* **12** 34–39.

NIETO-BARAJAS, L. E., MÜLLER, P., JI, Y., LU, Y. and MILLS, G. B. (2012). A Time-Series DDP for Functional Proteomics Profiles. *Biometrics* **68** 859–868.

PAGE, G. L., QUINTANA, F. A. and DAHL, D. B. (2021). Dependent Modeling of Temporal Sequences of Random Partitions. *Journal of Computational and Graphical Statistics* 1–14.

PATI, D., DUNSON, D. B. and TOKDAR, S. T. (2013). Posterior consistency in conditional distribution estimation. *Journal of Multivariate Analysis* **116** 456–472.

RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2010). Latent stick-breaking processes. *Journal of the American Statistical Association* **105** 647–659.

RODRIGUEZ, A. and DUNSON, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian analysis* **6**.

RODRIGUEZ, A. and TER HORST, E. (2008). Bayesian dynamic density estimation. *Bayesian Analysis* **3** 339–365.

ROSEN, R. (2013). *The world split open: How the modern women's movement changed America*. Tantor eBooks.

SCRUCCA, L., FOP, M., MURPHY, T. B. and RAFTERY, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal* **8** 289–317.

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

SIX, B. and ECKES, T. (1991). A closer look at the complex structure of gender stereotypes. *Sex roles* **24** 57–71.

TADDY, M. A. (2010). Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime. *Journal of the American Statistical Association* **105** 1403–1417.

WADE, S. and GHAHRAMANI, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis* **13** 559–626.

WILLIAMS, J. E. and BEST, D. L. (1990). *Measuring sex stereotypes: a multination study*. age Publications, Thousand Oaks, CA.

XIAO, S., KOTTAS, A. and SANSÓ, B. (2015). Modeling for seasonal marked point processes: An analysis of evolving hurricane occurrences. *The Annals of Applied Statistics* **9** 353–382.

ZHAO, J., WANG, T., YATSKAR, M., ORDONEZ, V. and CHANG, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen*.

Year 1900	**Cluster 1 (Neutral)** Mean (-0.011) Std (0.03) Min (-0.055) Max (0.087) Count (33)	**Cluster 2 (Man)** Mean (-0.094) Std (0.021) Min (-0.151) Max (-0.051) Count (43)
	architect, attendant, baker, blacksmith, broker, carpenter, cashier, chemist, clergy, clerical, conductor, cook, doctor, driver, gardener, housekeeper, instructor, janitor, laborer, mechanic, musician, nurse, operator, painter, pilot, porter, sailor, sales, scientist, surgeon, tailor, teacher, weaver	accountant, administrator, artist, athlete, auctioneer, author, bailiff, clerk, collector, dancer, dentist, designer, economist, engineer, farmer, geologist, guard, inspector, judge, lawyer, librarian, manager, mason, mathematician, midwife, official, photographer, physician, physicist, police, postmaster, professor, psychologist, retired, secretary, sheriff, shoemaker, smith, soldier, statistician, student, supervisor, surveyor
Year 1950	**Cluster 1 (Neutral)** Mean (-0.005) Std (0.02) Min (-0.048) Max (0.051) Count (38)	**Cluster 2 (Man)** Mean (-0.085) Std (0.02) Min (-0.136) Max (-0.043) Count (36)
	administrator, architect, athlete, attendant, auctioneer, baker, broker, carpenter, chemist, clergy, clerical, collector, conductor, cook, dancer, dentist, designer, economist, gardener, housekeeper, inspector, instructor, mason, mechanic, musician, operator, painter, photographer, physicist, porter, psychologist, sailor, scientist, supervisor, surgeon, tailor, teacher,	accountant, artist, author, bailiff, blacksmith, cashier, clerk, doctor, driver, engineer, farmer, geologist, guard, janitor, judge, laborer, lawyer, librarian, manager, mathematician, midwife, official, physician, pilot, police, postmaster, professor, retired, sales, secretary, sheriff, shoemaker, smith, soldier, statistician, surveyor
Year 2000	**Cluster 1 (Neutral)** Mean (-0.034) Std (0.04) Min (-0.163) Max (0.062) Count (76)	
	accountant, administrator, architect, artist, athlete, attendant, auctioneer, author, bailiff, baker, blacksmith, broker, carpenter, cashier, chemist, clergy, clerical, clerk, collector, conductor, cook, dancer, dentist, designer, doctor, driver, economist, engineer, farmer, gardener, geologist, guard, housekeeper, inspector, instructor, janitor, judge, laborer, lawyer, librarian, manager, mason, mathematician, mechanic, midwife, musician, nurse, official, operator, painter, photographer, physician, physicist, pilot, police, porter, postmaster, professor, psychologist, retired, sailor, sales, scientist, secretary, sheriff, shoemaker, smith, soldier, statistician, student, supervisor, surgeon, surveyor, tailor, teacher, weaver	

TABLE 2

Cluster estimates of the occupational bias data for $t = 1900, 1950, 2000$ *without singletons.*

	Cluster 1 (Woman) Mean (0.027) Std (0.022) Min (-0.023) Max (0.096) Count (102)	**Cluster 2 (Man)** Mean (-0.088) Std (0.004) Min (-0.09) Max (-0.069) Count (70)	**Cluster 3 (Man)** Mean (-0.018) Std (0.017) Min (-0.053) Max (0.017) Count (41)	**Cluster 4 (Woman)** Mean (0.044) Std (0.025) Min (0.004) Max (0.074) Count (8)	**Cluster 5 (Neutral)** Mean (0.021) Std (0.021) Min (0.003) Max (0.056) Count (4)	**Cluster 6 (Man)** Mean (-0.05) Std (0.009) Min (-0.062) Max (-0.041) Count (3)	**Cluster 7 (Woman)** Mean (0.106) Std (0.012) Min (0.093) Max (0.118) Count (2)
Year 1900	aggressive, alert, aloof, ambitious, artistic, attractive, awkward, calm, careless, cautious, cheerful, clever, coarse, commonplace, complaining, complicated, confused, considerate, contented, conventional, courageous, cowardly, cruel, curious, cynical, defensive, deliberate, dependent, dignified, discreet, dreamy, dull, efficient, emotional, energetic, fearful, foolish, friendly, frivolous, gentle, hasty, healthy, helpful, humorous, impatient, impulsive, indifferent, industrious, informal, initiative, jolly, kind, lazy, leisurely, logical, loyal, mild, mischievous, moderate, nervous, noisy, patient, peculiar, persistent, pleasant, polished, precise, progressive, queer, quick, quiet, rational, rebellious, restless, retiring, rigid, robust, rude, sensitive, sentimental, serious, shallow, silent, simple, sincere, sly, spontaneous, stubborn, superstitious, suspicious, tense, thoughtful, thrifty, timid, tough, trusting, understanding, unkind, unselfish, warm, wholesome, withdrawn	active, adaptable, appreciative, arrogant, assertive, autocratic, changeable, conceited, deceitful, dependable, distrustful, effeminate, enterprising, fickle, forceful, forgetful, forgiving, greedy, headstrong, honest, idealistic, immature, independent, individualistic, infantile, inhibited, intolerant, inventive, irresponsible, irritable, meek, methodical, obliging, obnoxious, optimistic, organized, outgoing, outspoken, painstaking, peaceable, persevering, pessimistic, praising, prejudiced, preoccupied, quarrelsome, quitting, realistic, reflective, resentful, resourceful, responsible, sarcastic, sociable, sophisticated, stable, submissive, tactful, talkative, thankless, tolerant, unaffected, unassuming, unfriendly, unstable, versatile, vindictive, wary, witty, worrying	adventurous, affected, anxious, bitter, capable, civilized, cold, confident, conscientious, conservative, cooperative, daring, determined, disorderly, dissatisfied, dominant, enthusiastic, generous, handsome, hostile, hurried, imaginative, ingenious, intelligent, loud, mature, moody, natural, practical, reckless, reliable, reserved, severe, shrewd, slow, steady, strong, thorough, unscrupulous, weak, wise	affectionate, formal, gloomy, modest, poised, selfish, shy, sympathetic	cool, demanding, masculine, relaxed	original, reasonable, stern	charming, feminine
	Cluster 1 (Neutral) Mean (0.009) Std (0.024) Min (-0.049) Max (0.059) Count (133)	**Cluster 2 (Man)** Mean (-0.081) Std (0.004) Min (-0.111) Max (-0.079) Count (73)	**Cluster 3 (Man)** Mean (-0.054) Std (0.014) Min (-0.08) Max (-0.032) Count (6)	**Cluster 4 (Woman)** Mean (0.044) Std (0.026) Min (0.006) Max (0.098) Count (7)	**Cluster 5 (Neutral)** Mean (0.019) Std (0.028) Min (-0.018) Max (0.052) Count (8)	**Cluster 6 (Woman)** Mean (0.118) Std (0.003) Min (0.115) Max (0.121) Count (2)	
Year 1950	affected, aggressive, alert, aloof, ambitious, anxious, arrogant, artistic, awkward, bitter, calm, capable, careless, cautious, cheerful, civilized, clever, cold, commonplace, complaining, complicated, confused, conscientious, contented, conventional, courageous, cruel, curious, cynical, daring, defensive, deliberate, demanding, dependent, determined, dignified, dissatisfied, dominant, efficient, emotional, energetic, enthusiastic, fearful, foolish, forceful, friendly, generous, gloomy, greedy, handsome, hasty, healthy, helpful, hostile, humorous, imaginative, impatient, indifferent, informal, ingenious, initiative, irresponsible, jolly, kind, lazy, leisurely, logical, loud, loyal, mature, mild, moderate, modest, moody, natural, nervous, noisy, optimistic, outgoing, outspoken, patient, peculiar, persistent, pleasant, poised, polished, practical, preoccupied, progressive, queer, quick, quiet, rational, realistic, reasonable, reckless, relaxed, reliable, reserved, restless, retiring, rigid, rude, selfish, sensitive, sentimental, serious, severe, shrewd, silent, simple, sincere, slow, sly, sophisticated, spontaneous, stable, stern, strong, stubborn, suspicious, sympathetic, tense, thorough, thoughtful, timid, tolerant, understanding, warm, weak, withdrawn, witty, worrying	adaptable, adventurous, appreciative, assertive, autocratic, changeable, conceited, confident, conservative, considerate, cowardly, deceitful, dependable, discreet, disorderly, distrustful, dreamy, enterprising, fickle, forgetful, forgiving, frivolous, headstrong, honest, idealistic, immature, impulsive, individualistic, industrious, infantile, inhibited, intolerant, inventive, irritable, meek, methodical, mischievous, obliging, obnoxious, organized, painstaking, peaceable, persevering, pessimistic, praising, prejudiced, quarrelsome, quitting, rebellious, reflective, resentful, resourceful, robust, sarcastic, sociable, submissive, superstitious, tactful, talkative, thankless, thrifty, tough, trusting, unaffected, unassuming, unfriendly, unkind, unscrupulous, unselfish, unstable, versatile, vindictive, wholesome	active, effeminate, independent, original, responsible, wise	attractive, dull, formal, hurried, masculine, precise, shy	affectionate, coarse, cool, cooperative, gentle, intelligent, steady, wary	charming, feminine	
	Cluster 1 (Neutral) Mean (0.003) Std (0.019) Min (-0.051) Max (0.076) Count (138)	**Cluster 2 (Man)** Mean (-0.051) Std (0.004) Min (-0.076) Max (-0.03) Count (77)	**Cluster 3 (Neutral)** Mean (0.034) Std (0.035) Min (-0.019) Max (0.08) Count (4)	**Cluster 4 (Neutral)** Mean (-0.008) Std (0.01) Min (-0.019) Max (0.011) Count (5)	**Cluster 5 (Man)** Mean (-0.026) Std (0.019) Min (-0.053) Max (-0.013) Count (3)		
Year 2000	active, affectionate, alert, ambitious, anxious, artistic, attractive, awkward, bitter, calm, careless, cautious, cheerful, civilized, clever, coarse, cold, commonplace, complaining, complicated, confused, cool, cooperative, courageous, cruel, curious, cynical, daring, deliberate, demanding, dependent, determined, dignified, discreet, dominant, dreamy, efficient, emotional, enthusiastic, fearful, foolish, formal, friendly, generous, gentle, gloomy, helpful, honest, hurried, imaginative, impatient, independent, indifferent, informal, initiative, intelligent, irresponsible, jolly, kind, lazy, leisurely, logical, loud, masculine, mature, mild, mischievous, moderate, modest, moody, natural, nervous, noisy, optimistic, organized, original, outgoing, outspoken, patient, peculiar, persistent, pleasant, poised, polished, practical, precise, preoccupied, progressive, quick, quiet, rational, realistic, reasonable, rebellious, reckless, reflective, relaxed, reliable, reserved, responsible, restless, rigid, robust, rude, sarcastic, selfish, sensitive, sentimental, serious, severe, shallow, shy, silent, simple, sincere, slow, sly, sophisticated, spontaneous, stable, steady, stern, strong, suspicious, sympathetic, tactful, tense, thorough, thoughtful, timid, trusting, understanding, unstable, versatile, wary, weak, withdrawn, worrying	adaptable, adventurous, aggressive, aloof, appreciative, assertive, autocratic, capable, changeable, conceited, conscientious, conservative, considerate, contented, cowardly, deceitful, dependable, disorderly, dissatisfied, distrustful, effeminate, enterprising, fickle, forceful, forgetful, forgiving, frivolous, handsome, hasty, headstrong, hostile, humorous, idealistic, immature, impulsive, individualistic, industrious, infantile, ingenious, inhibited, intolerant, inventive, irritable, meek, methodical, obliging, obnoxious, painstaking, peaceable, persevering, pessimistic, praising, prejudiced, quarrelsome, queer, quitting, resentful, resourceful, retiring, shrewd, sociable, submissive, superstitious, talkative, thankless, thrifty, tolerant, tough, unaffected, unassuming, unfriendly, unkind, unscrupulous, unselfish, vindictive, wholesome, wise	affected, dull, feminine, healthy	confident, energetic, greedy, stubborn, witty	arrogant, conventional, loyal		

TABLE 3

Cluster estimates of the adjective bias data for $t = 1900, 1950, 2000$ *without singletons.*

APPENDIX A: Data and further posterior inference for the gender stereotypes example - word emebeddings. All word lists for this application are from Garg et al. (2018), available on their GitHub page at `https://github.com/nikhgarg/EmbeddingDynamicStereotypes`. In this Appendix we explain how the two variables of interest, gender embedding bias referring to occupational and adjective lists, respectively, have been derived.

First of all, we consider four collated work lists from Garg et al. (2018) to represent each gender (men, women) and neutral words (occupations and adjectives), which we denote here by W_{man}, W_{woman}, W_{occu} and W_{adj}. The lists follow here:

- *Man words* (W_{man}): he, son, his, him, father, man, boy, himself, male, brother, sons, fathers, men, boys, males, brothers, uncle, uncles, nephew, nephews.
- *Woman words* (W_{woman}): she, daughter, hers, her, mother, woman, girl, herself, female, sister, daughters, mothers, women, girls, femen, sisters, aunt, aunts, niece, nieces.
- *Occupations* (W_{occu}): janitor, statistician, midwife, bailiff, auctioneer, photographer, geologist, shoemaker, athlete, cashier, dancer, housekeeper, accountant, physicist, gardener, dentist, weaver, blacksmith, psychologist, supervisor, mathematician, surveyor, tailor, designer, economist, mechanic, laborer, postmaster, broker, chemist, librarian, attendant, clerical, musician, porter, scientist, carpenter, sailor, instructor, sheriff, pilot, inspector, mason, baker, administrator, architect, collector, operator, surgeon, driver, painter, conductor, nurse, cook, engineer, retired, sales, lawyer, clergy, physician, farmer, clerk, manager, guard, artist, smith, official, police, doctor, professor, student, judge, teacher, author, secretary, soldier.
- *Adjectives* (W_{adj}): headstrong, thankless, tactful, distrustful, quarrelsome, effeminate, fickle, talkative, dependable, resentful, sarcastic, unassuming, changeable, resourceful, persevering, forgiving, assertive, individualistic, vindictive, sophisticated, deceitful, impulsive, sociable, methodical, idealistic, thrifty, outgoing, intolerant, autocratic, conceited, inventive, dreamy, appreciative, forgetful, forceful, submissive, pessimistic, versatile, adaptable, reflective, inhibited, outspoken, quitting, unselfish, immature, painstaking, leisurely, infantile, sly, praising, cynical, irresponsible, arrogant, obliging, unkind, wary, greedy, obnoxious, irritable, discreet, frivolous, cowardly, rebellious, adventurous, enterprising, unscrupulous, poised, moody, unfriendly, optimistic, disorderly, peaceable, considerate, humorous, worrying, preoccupied, trusting, mischievous, robust, superstitious, noisy, tolerant, realistic, masculine, witty, informal, prejudiced, reckless, jolly, courageous, meek, stubborn, aloof, sentimental, complaining, unaffected, cooperative, unstable, feminine, timid, retiring, relaxed, imaginative, shrewd, conscientious, industrious, hasty, commonplace, lazy, gloomy, thoughtful, dignified, wholesome, affectionate, aggressive, awkward, energetic, tough, shy, queer, careless, restless, cautious, polished, tense, suspicious, dissatisfied, ingenious, fearful, daring, persistent, demanding, impatient, contented, selfish, rude, spontaneous, conventional, cheerful, enthusiastic, modest, ambitious, alert, defensive, mature, coarse, charming, clever, shallow, deliberate, stern, emotional, rigid, mild, cruel, artistic, hurried, sympathetic, dull, civilized, loyal, withdrawn, confident, indifferent, conservative, foolish, moderate, handsome, helpful, gentle, dominant, hostile, generous, reliable, sincere, precise, calm, healthy, attractive, progressive, confused, rational, stable, bitter, sensitive, initiative, loud, thorough, logical, intelligent, steady, formal, complicated, cool, curious, reserved, silent, honest, quick, friendly, efficient, pleasant, severe, peculiar, quiet, weak, anxious, nervous, warm, slow, dependent, wise, organized, affected, reasonable, capable, active, independent, patient, practical, serious, understanding, cold, responsible, simple, original, strong, determined, natural, kind.

For each word in those lists, we download (from https://nlp.stanford.edu/projects/histwords) the corresponding embeddings from previously trained Genre-Balanced American English

embeddings from Corpus of Historical American English (COHA) (Hamilton, Leskovec and Jurafsky, 2016) for the eleven decades available, specifically referring to decades 1900, 1910, ..., 1990, 2000. Following Garg et al. (2018), from the embeddings and word lists, we are able to measure the embedding bias (i.e. strength of association) between words that represent gender groups (i.e. women and men) and neutral words such as Occupations and Adjectives. Here subscript t represents the year (decade), i.e. $t = 1900, 1910, \ldots, 1990, 2000$.

- For each time t, we compute $\mathbf{m}_t$, the average embedding vector for W_{man}, averaged over the 20 words in the list W_{man}; similarly, $\mathbf{w}_t$ denotes the average embedding for W_{woman}, averaged over all the word embeddings in the list W_{woman}.
- We define the embedding bias Y'_{tj} for (or against) women of the jth occupation word in W_{occu} at time t as the difference of the Euclidean distances between the jth word and the men representative and the difference of the Euclidean distances between the jth word and the women average vector, i.e.

$$\text{occupation bias}_{tj} = Y'_{tj} = ||\mathbf{o}_{tj} - \mathbf{m}_t||_2 - ||\mathbf{o}_{tj} - \mathbf{w}_t||_2,$$

  where $\mathbf{o}_{tj}$ is the word embedding vector of the jth occupation word at time t and $||\cdot||_2$ is the Euclidean norm of a finite-dimensional vector. We standardize the occupation embedding bias for women by considering Y_{tj} as Y'_{tj} minus the overall mean and then divide it by the overall standard deviation.
- Similarly, we define the adjective embedding bias Z'_{tl} for (or against) women of the lth adjective word in W_{adj} at time t, as the difference

$$\text{adjective bias}_{tl} = Z'_{tl} = ||\mathbf{a}_{tl} - \mathbf{m}_t||_2 - ||\mathbf{a}_{tl} - \mathbf{w}_t||_2,$$

  where $\mathbf{a}_{tl}$ is the word embedding vector of the lth adjective word at time t. Here each Z'_{tl} has been standardized subtracting from Z'_{tl} the overall mean and then divide the difference by the overall standard deviation to get Z_{tl} as for the occupational bias for women.

Note that, if the bias value is negative, then the embedding more closely associates the occupation (or adjective) word with men, because the distance between the occupation (or adjective) word is closer to men than women. Other norm definitions could be used here, as, for instance, cosin similarity. Hence, gender bias *against women* corresponds to negative values of the embedding bias.

There are many ties in both original datasets $\{Y'_{tj}\}$ and $\{Z'_{tj}\}$. There are different strategies to deal with ties in the data, such as opting for a truncated normal distribution as the kernel $k(\cdot;\theta)$ of the mixture. However, since our primary interest is to investigate the cluster estimates, we jitter the data by adding zero-mean Gaussian noise, with variance 0.001 to both bias data (i.e. "jittering") before standardization; the ratio between the noise variance and the overall range of data (max - min) is 0.3% in case of Occupations and 0.4% in case of Adjectives.

It is clear from Figure 12 that most of the occupation words show gender bias against women, since the empirical distributions as shown by the boxplots are all concentrated below 0. With respect to the adjective bias, from Figure 13 we see that the adjectives are biased against women too, but less than the occupation words.

As exploratory data analysis, we compute data-driven cluster estimates of occupation and adjective embedding bias through the R packages *cluster* and *factoextra*. Typical methods are based on the distance between pairs of embeddings. We show the Euclidean distance between pairs of occupation embedding bias for each decade in Figure 14 through the command `fvi_dist` of package *factoextra*. To make a comparison over time, for each decade we fix the (alphabetical) order of the occupation words on the axes. Even though it is difficult to highlight particular changes over time in the matrices in Figure 14, it seems clear that

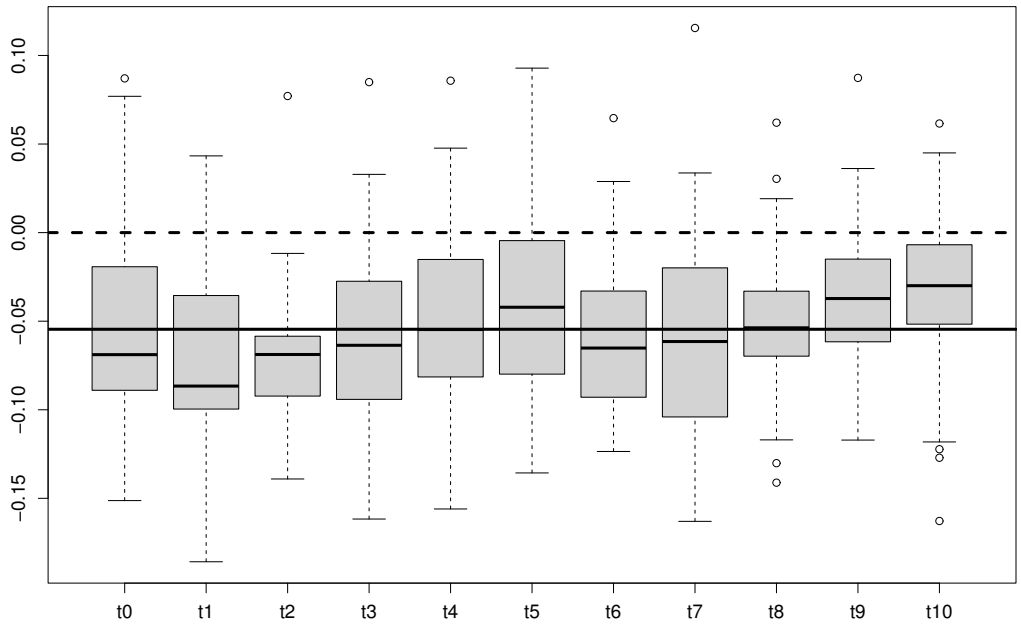

Fig 12: Boxplots of the occupation embedding bias across decades. The horizontal black line indicates the overall mean and the dashed line corresponds to zero.

the distances between words to not change abruptly from one decade to the other. Moreover, we compute data-driven cluster estimates of occupation and adjective embedding bias using K-means and Partitioning Around Medoids (PAM) algorithms. The elbow method (plots not shown here) for K-means indicates that the optimal number of clusters is constant through decades and it is equal to four and three for the occupation and adjective embedding bias, respectively. However, because of the sensitivity of K-means cluster estimates to initial random allocation, to the order of data and to outliers, we estimate cluster allocation using PAM as well. PAM (Kaufman and Rousseeuw, 1990) is an iterative clustering methods based on k-medoids, alternative to K-means centroids. For each decade, the optimal number of clusters is determined as the integer k (between 2 and 10) that maximizes the average silhouette index as implemented in the R packages *cluster* and *factoextra*. The average silhouette is one of the many internal cluster validation criteria, measuring the quality of a clustering. A high average silhouette width (close to 1) corresponds to a satisfactory clustering structure. See Kassambara (2017) for further details.

Boxplots of occupation embedding bias per estimated clusters (via PAM) across decades are displayed in Figure 15. ~~Data points are in red, while the boxplots and outliers are represented in black.~~ Note that data points which are also outliers are reported twice. The number of boxplots is the number of clusters in each optimal partition. A similar plot for the adjective embedding bias is displayed in Figure 16.

Tables 4 and 5 show the sample sizes of the three largest estimated clusters (via PAM) for the occupation and adjective embedding bias, respectively, for each decade. These sample sizes do not generally show abrupt changes over time.

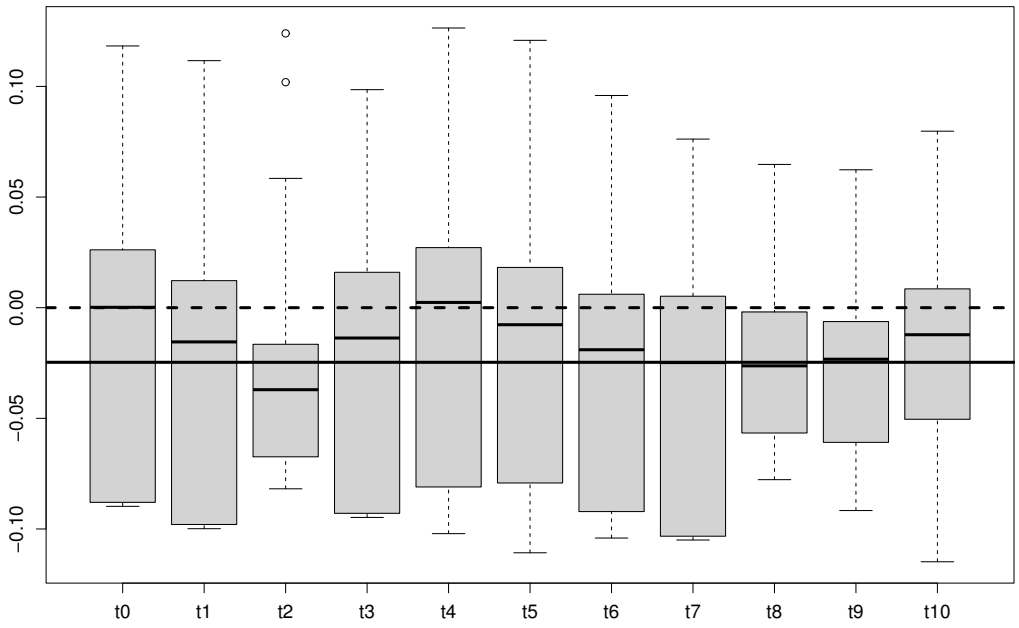

Fig 13: Boxplots of the adjective embedding bias across decades. The horizontal black line indicates the overall mean and the dashed line corresponds to zero.

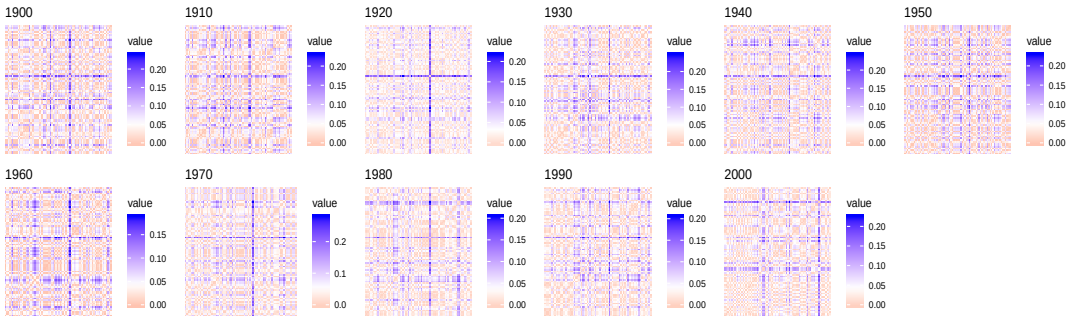

Fig 14: Distance matrices of occupation embedding bias across decades.

decade	**1900**	**1910**	**1920**	**1930**	**1940**	**1950**
n_1, n_2, n_3	43, 33, -	26,9,8	24,19,9	18,10,9	39, 37, -	38, 38, -
Average silhouette	0.690	0.701	0.670	0.653	0.673	0.696
decade	**1960**	**1970**	**1980**	**1990**	**2000**	
n_1, n_2, n_3	39, 37, -,	21, 15, 9	22, 17, 10	22, 17, 13	17,14,13	
Average silhouette	0.678	0.685	0.628	0.632	0.676	

TABLE 4

Sample sizes of the three largest estimated clusters and average silhouette index (via PAM) for the occupation embedding bias for each decade.

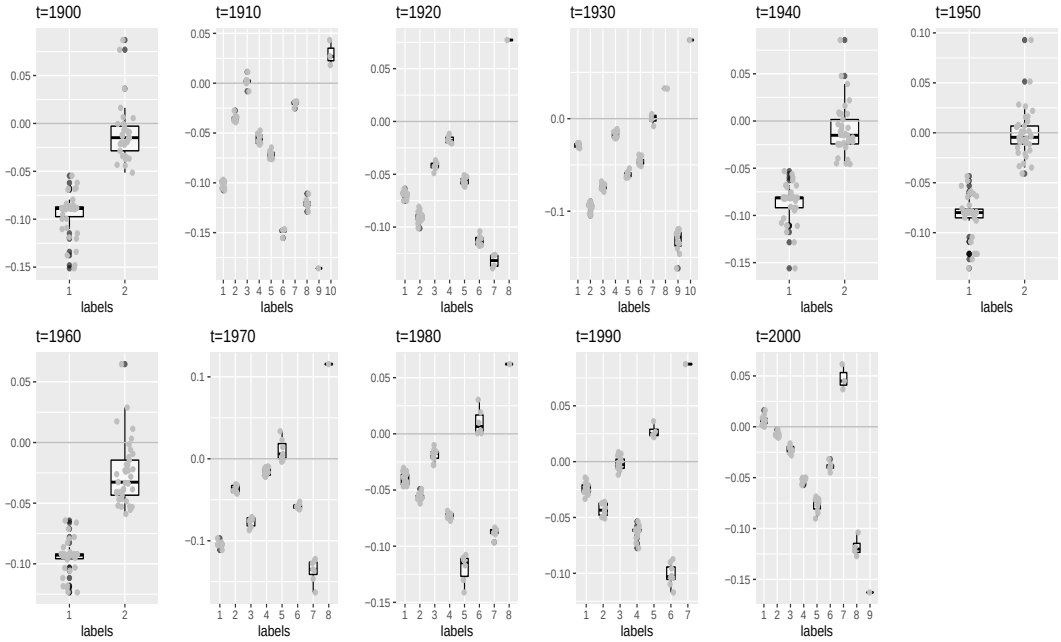

Fig 15: Boxplots of occupation embedding bias per estimated clusters (via PAM) across decades. Data points are in grey, while the boxplots and outliers are represented in black.

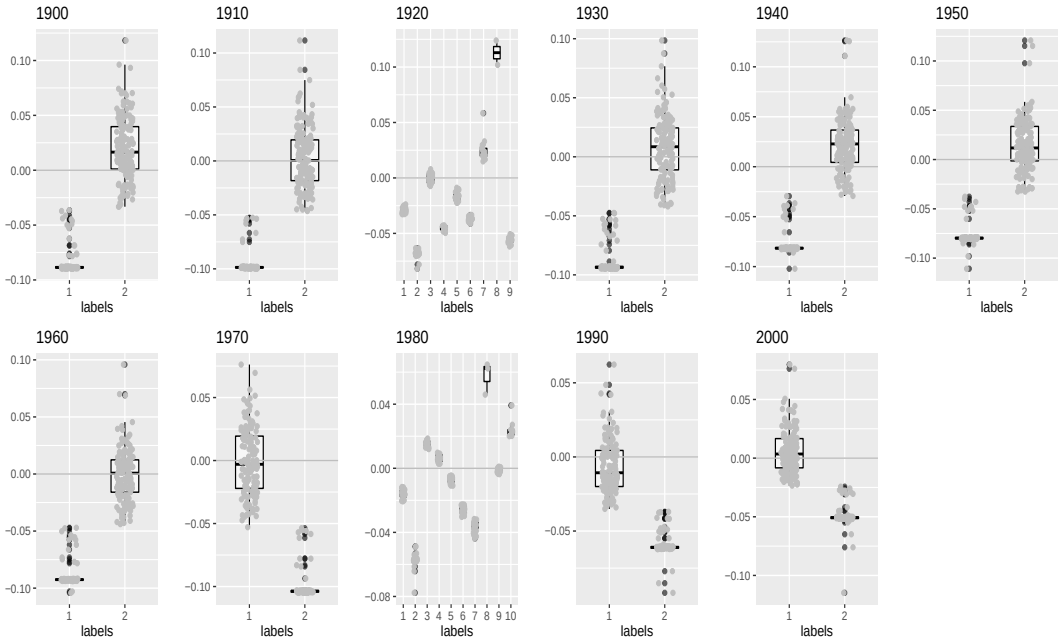

Fig 16: Boxplots of adjective embedding bias per estimated clusters (via PAM) across decades. Datapoints are in grey, while the boxplots and outliers are represented in black.

From this brief exploratory data analysis, we believe that we can conclude that an auto-regressive behaviour in the clustering structure over time is a reasonable assumption. Both methods, K-means and PAM, suggest that reasonable values for the number K of clusters for both datasets vary between two and ten (taken to be the largest possible value in this

decade	**1900**	**1910**	**1920**	**1930**	**1940**	**1950**
n_1, n_2, n_3	150, 80, -	159,71, -	66,31,27	151, 79, -	148, 81, -	146, 84, -
Average silhouette	0.727	0.720	0.667	0.721	0.747	0.741
decade	**1960**	**1970**	**1980**	**1990**	**2000**	
n_1, n_2, n_3	145, 85, -	147, 83, -	76,27,23	134,96,-	141,89, -	
Average silhouette	0.746	0.744	0.676	0.710	0.684	

TABLE 5

Sample sizes of the three largest estimated clusters and average silhouette index (via PAM) for the adjective embedding bias for each decade.

exploratory analysis). As such, we specify the prior on the concentration parameter M in the analysis of Section 3 to induce a prior mean on the number of clusters equal to four.

In this Appendix we also include extra figures of the gender bias examples. The right panel of Figure 17 shows the marginal posterior distribution of ψ, with posterior mean equal to 0.072, in the case of the adjective embedding bias data in Section 3.2. The posterior co-clustering probabilities in this case are displayed in Figure 18. ~~The predominant colors are red, green and blue, and the left and middle panel (1900 and 1950) are very similar. Instead the right panel (2000) has very little green and most blue (low co-clustering probability) and red (high co-clustering probability).~~ The three co-clustering plots seem similar. *La frase che c'era prima era sbagliata se la Figure 18 è esatta.*

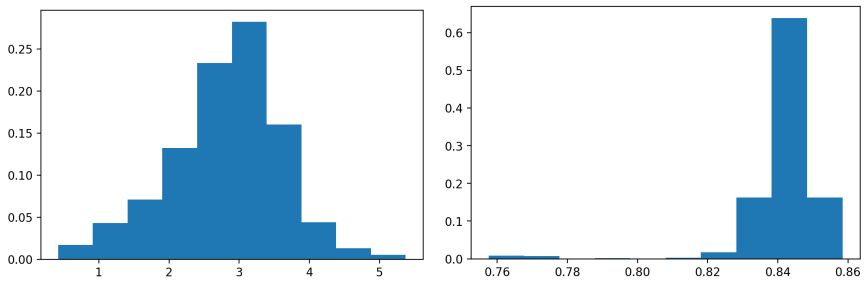

Fig 17: Marginal posterior distributions of M (left panel) and ψ (right panel) for adjectives bias data.

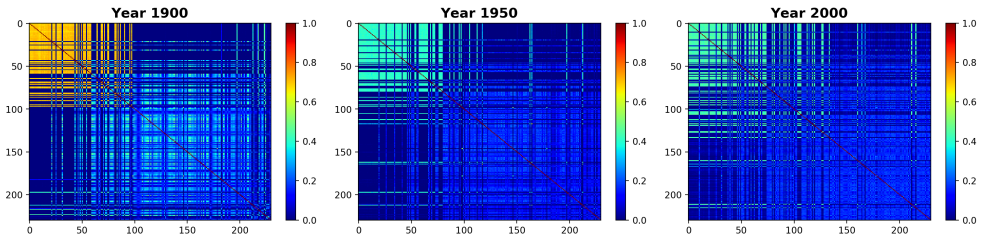

Fig 18: Posterior co-clustering for adjectives bias data for $t = 1990$ (left panel), $t = 1950$ (center panel) and $t = 2000$ (right panel).