arTs Università degli Studi di Trieste
Archivio della ricerca – postprint

Bayesian semiparametric modelling of contraceptive behaviour in India via sequential logistic regressions

Tommaso Rigon,

Bocconi University, Milan, Italy

Daniele Durante

Bocconi University and Bocconi Institute for Data Science, Milan, Italy

and Nicola Torelli

University of Trieste, Italy

Summary. Family planning has been characterized by highly different strategic programmes in India, including method-specific contraceptive targets, coercive sterilization and more recent target-free approaches. These major changes in family planning policies over time have motivated considerable interest towards assessing the effectiveness of the different planning programmes. Current studies mainly focus on the factors driving the choice among specific subsets of contraceptives, such as a preference for alternative methods other than sterilization. Although this restricted focus produces key insights, it fails to provide a global overview of the different policies, and of the determinants underlying the choices from the entire range of contraceptive methods. Motivated by this consideration, we propose a Bayesian semiparametric model relying on a reparameterization of the multinomial probability mass function via a set of conditional Bernoulli choices. This binary decision tree is defined to be consistent with the current family planning policies in India, and coherent with a reasonable process characterizing the choice between increasingly nested subsets of contraceptive methods. The model allows a subset of covariates to enter the predictor via Bayesian penalized splines and exploits mixture models to represent uncertainty in the distribution of the state-specific random effects flexibly. This combination of flexible and careful reparameterizations allows a broader and interpretable overview of the policies and contraceptive preferences in India.

Keywords: Bayesian inference; Contraceptive method; Mixture model; Penalized splines; Pólya–gamma schemes; Sequential logistic regression

1. Introduction

The rapid population growth in developing countries is a key topic in demographic research, having immediate consequences on the increasing demand for social services, rising rates of unemployment and reduced standard of living. Although there is still a debated literature concerning the long-term effects of overpopulation on socio-economic growth (e.g. Bloom (2012)), in the view of the developing countries, limited resources and the rapid population growth are important barriers for the short-term development process, and a main concern for governments' policies. This is particularly true in India, where the coexistence of early marriages, high

Address for correspondence: Tommaso Rigon, Department of Decision Sciences, Bocconi University, Via Roentgen 1, Milan 20136, Italy.
E-mail: tommaso.rigon@phd.unibocconi.it

rates of poverty, illiteracy and a decline in infant mortality, favoured a population growth rate of 1.4%—double China's 0.7% (Bloom, 2012)—leading to an unsustainable population which is expected to reach 1.4 billion over the next quarter century.

Although India was the first nation to introduce an official family planning programme in 1951 (Pachauri, 2004, 2014), the broader access to welfare services under the clinic-based approach during the first and second 5-year plans in the 1950s, the subsequent focus on method-specific contraceptive targets in the mid-1960s and the coercive sterilization programmes in the 1970s and early 1980s failed to control the rapid population growth properly. Target-free contraceptive services, accounting for the different reproductive health needs of the population, were later promoted after the 1994 International Conference on Population and Developments (Pachauri, 2004, 2014). However, scepticism remains about the effectiveness of such services in increasing contraceptive prevalence, and in stimulating broader access to modern temporary methods, different from sterilization (e.g. Säävälä (1999)). In fact, there is evidence that a preference for permanent contraceptive practices is still dominant compared with reversible methods, and that most of the services that are provided by the public programmes relate to sterilization (Pachauri, 2004, 2014). Conversely, there is a growing effort by the private sector aimed at providing reproductive health services that are associated with reversible contraceptive methods (Pachauri, 2004, 2014). Refer to Harkavy and Roy (2007) and Chaurasia and Singh (2014) for additional details and timelines of family planning programmes in India.

The above differences in family planning services, combined with the marked sociodemographic inequalities characterizing the population in India, have increased the emphasis on the availability of strategic data sets and statistical models to evaluate the family planning policies, and to identify which subsets of the population have not been properly addressed. As a result, detailed surveys, such as the India Human Development Survey II (IHDS II) and the National Family Health Survey, have been recently conducted, motivating an increasing interest on the determinants of contraceptive choice in the light of the current policies. In this contribution we explore the IHDS II survey data at `http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/36151`, to provide a flexible overview of the various policies and preferences, within a single model. The data set is described in detail below.

1.1. India Human Development Survey II

Contraceptives play a relevant role in family planning, and the widely accepted association between contraceptive prevalence rates and total fertility rates (e.g. Mauldin and Segal (1988)) drives extensive demographic research on developing countries. Consistent with these interests, we focus our analysis on the IHDS II 2011–2012 module devoted to ever-married women, which provides information about contraceptive choices, along with other sociodemographic variables.

The IHDS II 2011–2012 is a nationally representative multitopic survey conducted on 42 152 households over 33 states of India (Desai *et al.*, 2015). The survey arises from collaboration between the University of Maryland and the National Council of Applied Economic Research in New Delhi and is divided into different modules aimed at monitoring a wide range of socio-economic behaviours. Eligible units in our analysis are women aged 15–49 years who have been married at least once in their life, although their current marital status may not be married, and who were not pregnant at the moment of the interview. Recalling our research interests described earlier in this section, the response in our study is a qualitative variable having four mutually exclusive outcomes.

(a) No contraceptive: no contraceptive method is used.
(b) Sterilization: the woman or her partner underwent sterilization or hysterectomy.

(c) Natural methods: either withdrawal or periodic abstinence is chosen.
(d) Modern methods: the woman or her partner uses modern methods (e.g. oral pill or condoms).

A small set of women declared that they used other types of contraceptives. Since we do not have information about these alternative methods, we held them out from our analysis. Consistent with this choice, we also do not consider women for whom the information on contraceptive preference is not observed. This preliminary preprocessing provides a final sample size of $n = 30\,524$. Although finer classifications of contraceptive methods could be considered, the four categories that are listed above are those of main interest in family planning policies (e.g. Pachauri (2004)) and are commonly considered in statistical modelling of contraceptive behaviour in India (e.g. De Oliveira *et al*. (2014)).

In selecting the covariates of interest we leverage instead recent evidence from statistical analyses on contraceptive behaviour in India (e.g. De Oliveira *et al*. (2014)), and more general discussions on the relevant factors underlying contraceptive preferences, provided by social science studies (e.g. Pachauri (2004)). More specifically, we consider the AGE information, a binary variable AREA indicating whether the woman lives in urban or rural areas, a four-level RELIGION factor, encoding hindu, muslim, christian or other religions, a categorical variable EDUCATION classifying women according to no education, low education, intermediate education or high education and a grouping variable CHILD for women having no child, one child or more than one child. We also exploit the information on the STATE of residence to define the hierarchy in our model via state-specific effects. These quantities are also of interest for inference, in providing information on across-state differences in contraceptive preferences, after controlling for the sociodemographic covariates. Although we could consider additional covariates, our main goal is to disambiguate and interpret the effect of the most studied variables at the various steps of the contraceptive choices.

As discussed in Section 2.1, we propose to analyse the above data under a statistical model which relies on a set of sequential binary comparisons among subsets of contraceptive choices. This focus is explicitly motivated by the current family planning policies in India that were discussed earlier and is also coherent with a reasonable decision process underlying contraceptive preferences, thereby providing more general and interpretable inference compared with classical analyses based on standard parameterizations of the multinomial logistic regression. Motivated by the marked sociodemographic differences characterizing the population in India, we also improve flexibility in modelling the covariates effects at the various steps of the sequential binary choices under a Bayesian semiparametric formulation outlined in Section 2.2. Beside facilitating flexible and interpretable inference, the methods proposed are also associated with simple algorithms for inference; see Section 3. As carefully outlined in Section 4, this combination of flexible representations for the covariates effects, and careful reparameterizations of the multinomial likelihood for the contraceptive preference data, provides relevant and interpretable insights for policy makers, while improving predictive performance. These results, and the source code to reproduce them, are available from `https://github.com/tommasorigon/India-SequentiaLogit` along with an interactive Shiny application. Concluding remarks are provided in Section 5.

2. Bayesian semiparametric modelling of contraceptive preferences

As discussed in Section 1, we reparameterize the multinomial probability mass function for the contraceptive methods via conditional Bernoulli choices for subsets of contraceptives and

provide inference on the sociodemographic factors underlying these sequential binary comparisons via a Bayesian semiparametric representation for the covariates effects. Sections 2.1 and 2.2 describe these generalizations, with a specific reference to the research interests that are associated with contraceptive behaviour in India.

2.1. Model formulation via sequential Bernoulli choices

Let $\mathbf{y}_{ij} = (y_{ij1}, y_{ij2}, y_{ij3}, y_{ij4})$ denote the vector of binary variables encoding the contraceptive choice of woman $j = 1, \ldots, n_i$ in state $i = 1, \ldots, 33$, with

(a) $\mathbf{y}_{ij} = (1, 0, 0, 0)$ if woman j in state i and her partner use no contraceptive methods,
(b) $\mathbf{y}_{ij} = (0, 1, 0, 0)$ if woman j in state i or her partner underwent sterilization,
(c) $\mathbf{y}_{ij} = (0, 0, 1, 0)$ if woman j in state i and her partner use natural methods and
(d) $\mathbf{y}_{ij} = (0, 0, 0, 1)$ if woman j in state i or her partner uses modern methods.

Following standard procedures in statistical modelling of categorical response data we let

$$\mathbf{y}_{ij} | \boldsymbol{\pi}_{ij} \sim \text{Multinom}(1, \boldsymbol{\pi}_{ij}), \qquad \boldsymbol{\pi}_{ij} = (\pi_{ij1}, \pi_{ij2}, \pi_{ij3}, \pi_{ij4}), \tag{1}$$

independently for each state $i = 1, \ldots, 33$ and woman $j = 1, \ldots, n_i$, where $\pi_{ijr} \in (0, 1)$ indicates the probability that woman j in state i adopts contraceptive behaviour r, listed in Section 1.1.

Under the usual specification of multinomial regression (e.g. Agresti (2007)), we could pair each contraceptive choice with a reference category—characterizing, for example, no use of contraceptives—and then model the log-odds of every pair as a function of the covariates. Although this is a common specification (e.g. Jayaraman *et al.* (2009), Husain *et al.* (2013) and Ram *et al.* (2014)), the resulting inference and conclusions are confined to pairwise comparisons with the selected reference category, thereby ruling out the possibility to learn—within a single statistical model—the direct effect of the covariates on other log-odds of potential interest for policy makers, including conditional choices between subsets of similar contraceptive methods. Indeed, these conditional quantities are typically of interest and are implicitly involved in the statistical analysis of contraceptive behaviour. For example, De Oliveira *et al.* (2014) relied on a multinomial logistic regression, applied to the subset of women who were currently using contraceptives, to disentangle the sociodemographic factors that are associated with the preference of modern and traditional contraceptive methods compared with sterilization. Mishra *et al.* (2014) considered instead different logistic regressions to learn the covariates effects on the preference of a specific contraceptive method—or a subset of methods—for different combinations of interest.

The above contributions arguably rely on more interpretable parameterizations and comparisons, which can be reformulated depending on the research interests. For instance, when the main goal is to evaluate policies that are aimed at increasing contraceptive prevalence under the current target-free 'cafeteria' approach in India (Pachauri, 2004, 2014), it is arguably more coherent and of direct interest to study the sociodemographic factors underlying the binary decision to use or not contraceptives, rather than modelling the log-odds of each contraceptive method with no contraceptive usage as the reference category.

Motivated by these considerations, we rely on a reparameterization of the multinomial probability mass function via a sequence of Bernoulli choices between increasingly nested subsets of contraceptive methods. The binary tree structure of interest is represented in Fig. 1 and is defined to characterize a reasonable decision process underlying the contraceptive choices. In particular, as shown in Fig. 1, this decision process starts with the choice of using or not contraceptives. If a contraceptive method is chosen, the next step requires deciding between permanent

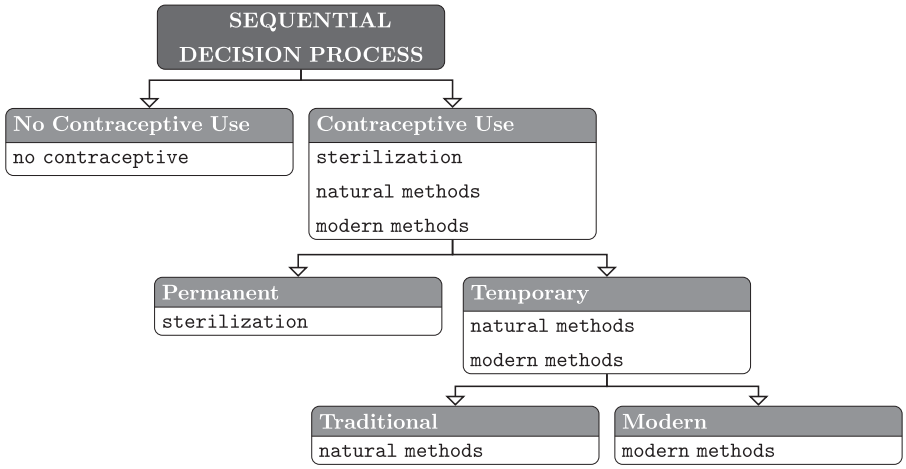

Fig. 1. Representation of the sequential process underlying contraceptive behaviour

or temporary contraceptives. Finally—in case a temporary method is preferred—the choice is between traditional or modern methods. Our overarching focus is to infer the sociodemographic effects underlying each step of this decision process.

Although other nested decision processes can be devised, and our inference procedures can be easily adapted to different binary tree structures, the sequential mechanism in Fig. 1 allows interpretable inference on dependence structures of direct interest in the light of the current India family programmes, within a single statistical model. In particular, disentangling—as a first step—the sociodemographic factors that are associated with the contraceptive prevalence is coherent with the current 'cafeteria' approach in India (Pachauri, 2004, 2014) and can provide relevant insights into which subsets of the population in India have not been currently addressed. Indeed, the contraceptive prevalence rate is one of the main performance indicators for family planning (e.g. Alkema *et al*. (2013)), and several analyses focus directly on this binary comparison, pooling the different contraceptive methods (e.g. McNay *et al*. (2003), Dharmalingam and Morgan (2004), De Oliveira and Dias (2014) and Haque and Patel (2015)).

The second step recalls, instead, the model that was proposed by De Oliveira *et al*. (2014) and is motivated by the fundamental interest underlying the factors that are associated with the ongoing dominance of sterilization compared with alternative methods promoted by target-free programmes in India. Consistent with De Oliveira *et al*. (2014), this analysis focuses on the subsets of women currently using a contraceptive method, to isolate the inference from the effects of use *versus* non-use of contraceptives. However, differently from De Oliveira *et al*. (2014)—who considered a multinomial logistic regression for sterilization, traditional and modern methods—we first study the choice between sterilization and the alternative temporary methods, and then focus on the decision between traditional and modern methods, for the subset of non-sterilized women. This arguably allows more direct inference—in the second step—on the sociodemographic factors underlying the general preference for sterilization, which is of main interest when the goal is to understand the reasons for the failure of family planning policies in motivating broader access to temporary methods (e.g. Säävälä (1999) and Pachauri (2004, 2014)).

Finally, as discussed in Section 1, there is a growing effort by the private sector in India towards addressing reproductive health needs other than sterilization, with a focus on modern reversible methods (e.g. Pachauri (2004, 2014)). Consistent with this, the last step in Fig. 1 focuses on the choice between natural and modern methods, for those women currently using

reversible contraceptives. This group arguably represents the segment of more direct interest for the private sector (e.g. Pachauri (2004, 2014)), and learning the determinants underlying the preference for natural methods instead of modern methods can provide key insights for the private sector to assess and improve targeting strategies.

As a consequence of the above sequential process, the explicit focus of inference is not directly on the vector of marginal probabilities $\boldsymbol{\pi}_{ij} = (\pi_{ij1}, \pi_{ij2}, \pi_{ij3}, \pi_{ij4})$ for the different contraceptive behaviours, but on the conditional probabilities $\boldsymbol{\rho}_{ij} = (\rho_{ij1}, \rho_{ij2}, \rho_{ij3}, \rho_{ij4} = 1 - \rho_{ij3})$, defined as

$$\rho_{ij1} = \sum_{r=2}^{4} \pi_{ijr}, \qquad \rho_{ij2} = \sum_{r=3}^{4} \pi_{ijr} \Big/ \sum_{r=2}^{4} \pi_{ijr}, \qquad \rho_{ij3} = \pi_{ij4} \Big/ \sum_{r=3}^{4} \pi_{ijr}, \tag{2}$$

for every state $i = 1, \ldots, 33$ and woman $j = 1, \ldots, n_i$. This reparameterization facilitates inference on the conditional probabilities characterizing the sequential process in Fig. 1. In fact, ρ_{ij1} represents the probability to use contraceptives, whereas ρ_{ij2} denotes the conditional probability of considering a reversible method, given the decision to use any contraceptive. Finally—following the sequential decision process in Fig. 1—the parameters ρ_{ij3} and ρ_{ij4} measure the probability of using a modern or traditional contraceptive method respectively, conditionally on the adoption of temporary methods.

Differently from De Oliveira *et al.* (2014), equation (2) allows direct modelling of the entire range of contraceptive behaviours, instead of just a subset of them, thereby providing more general inference and prediction, which are of interest for a wider spectrum of family planning policies. Moreover, inference is performed within a single statistical model based on a reparameterization of the multinomial probability mass function, providing coherent methods for prediction and quantification of uncertainty. This is not true in Mishra *et al.* (2014) whose conditional inference on combinations of contraceptive methods cannot be recast within a single statistical model. Conversely, under the proposed reparameterization in equation (2), the multinomial probability mass function $\text{pr}(\mathbf{y}_{ij}) = \pi_{ij1}^{y_{ij1}} \pi_{ij2}^{y_{ij2}} \pi_{ij3}^{y_{ij3}} \pi_{ij4}^{y_{ij4}}$ for each statistical unit can be formally rewritten as the product of Bernoulli probability mass functions for the sequential binary comparisons in Fig. 1, obtaining

$$\text{pr}(\mathbf{y}_{ij}) = \rho_{ij1}^{y_{ij2}+y_{ij3}+y_{ij4}} (1-\rho_{ij1})^{y_{ij1}} \rho_{ij2}^{y_{ij3}+y_{ij4}} (1-\rho_{ij2})^{y_{ij2}} \rho_{ij3}^{y_{ij4}} (1-\rho_{ij3})^{y_{ij3}}, \tag{3}$$

where the first Bernoulli variable characterizes the choice between no use ($y_{ij1} = 1; y_{ij2} + y_{ij3} + y_{ij4} = 0$) and use ($y_{ij1} = 0; y_{ij2} + y_{ij3} + y_{ij4} = 1$) of contraceptives, whereas the second represents the decision between sterilization ($y_{ij2} = 1; y_{ij3} + y_{ij4} = 0$) and reversible methods ($y_{ij2} = 0; y_{ij3} + y_{ij4} = 1$), for those women using contraceptives. Finally, the third Bernoulli variable focuses on the choice between natural ($y_{ij3} = 1; y_{ij4} = 0$) and modern methods ($y_{ij3} = 0; y_{ij4} = 1$), for the women currently using temporary contraceptives. Refer to Tutz (1991) for an overview of sequential logistic regression.

Consistent with our goals, and motivated by results in equation (3), we aim to learn the sociodemographic factors underlying the sequential binary choices in Fig. 1, via a logistic regression for each conditional probability in equation (2). To accomplish this goal accurately, we seek a flexible representation for the relationship between the conditional probabilities and the covariates, which allows coherent uncertainty quantification, efficient inference and possible inclusion of prior information. Consistent with these aims, we rely on a Bayesian semiparametric approach, which provides an appealing direction in hierarchical demographic models characterized by nested structures—such as those outlined in Fig. 1—and by the need to borrow information across the observed data to improve inference on parameters for which limited information is available. As discussed in Section 2.2, this is particularly useful in flexibly mod-

elling the functional effect of age, and the changes in contraceptive preferences across states having limited data. Refer to Bijak and Bryant (2016) and Elderd and Miller (2016) for a careful discussion on the benefits of Bayesian demographic inference in characterizing hierarchical and complex representations, quantifying and propagating uncertainty, borrowing of information and incorporating possible prior knowledge. Besides these benefits, there are also practical computational advantages, allowing simple posterior inference for the parameters of interest, covering the covariates effects, along with the conditional and the marginal probabilities in equations (2) and (1) respectively.

2.2. Bayesian semiparametric logistic regressions

Recalling the discussion in Section 2.1, the main focus is on learning the effects of the covariates on the conditional probabilities in equation (2), via the set of logistic regressions having additive effects

$$\text{logit}(\rho_{ijk}) = \log\left(\frac{\rho_{ijk}}{1-\rho_{ijk}}\right) = \mu_{ik} + f_k(\text{age}_{ij}) + \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_k, \qquad k = 1, \ldots, 3, \tag{4}$$

where $\mathbf{x}_{ij}$ is the vector of covariates encoding the variables AREA, RELIGION, EDUCATION and CHILD into dummy indicators, with associated vector of coefficients $\boldsymbol{\beta}_k$. Since the model intercept is incorporated in the functional effect of variable AGE, one category for each qualitative covariate is left out in equation (4) and considered as baseline to avoid the usual identifiability issues that are caused by the dummy trap in regression models. Hence, $\boldsymbol{\beta}_k$ represents incremental effects with respect to the baseline categories. Representation (4) consists of a state-specific coefficient μ_{ik}, a functional effect of the variable AGE and a set of dummy variables to learn changes in the conditional probabilities of interest with the categories of the qualitative variables AREA, RELIGION, EDUCATION and CHILD. Consistent with the above discussion, also the first state-specific effect μ_{1k}, corresponding to the most populated state in our sample—Uttar Pradesh—has been fixed to 0, and therefore $\mu_{2k}, \ldots, \mu_{33k}$ measure incremental effects with respect to this state. Although identifiability could be incorporated also via restrictions on the prior or via post-processing (e.g. Li *et al.* (2011)), we prefer to enforce identifiability directly in the likelihood to facilitate interpretation, and possible implementation in frequentist settings.

To be more specific, the three logistic regressions of interest are

$$\left.\begin{aligned} \text{pr}(\text{contraceptive use}) &= \rho_{ij1}, & \text{logit}(\rho_{ij1}) &= \mu_{i1} + f_1(\text{age}_{ij}) + \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1, \\ \text{pr}(\text{temporary} \mid \text{contraceptive use}) &= \rho_{ij2}, & \text{logit}(\rho_{ij2}) &= \mu_{i2} + f_2(\text{age}_{ij}) + \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_2, \\ \text{pr}(\text{modern} \mid \text{temporary}) &= \rho_{ij3}, & \text{logit}(\rho_{ij3}) &= \mu_{i3} + f_3(\text{age}_{ij}) + \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_3, \end{aligned}\right\} \tag{5}$$

and our focus is on providing flexible Bayesian inference on the parameters in equations (5). We consider the logistic link, instead of other alternatives—such as the probit link—since it provides a common and more interpretable choice in these types of analyses, while representing the canonical link in the exponential family representation of the Bernoulli random variable.

In modelling the state effects μ_{ik}, $i = 2, \ldots, 33$, $k = 1, \ldots, 3$—or similar community level covariates—current studies consider either fixed parameters within a classical generalized linear model framework (e.g. Jayaraman *et al.* (2009), Rai and Unisa (2013) and Ram *et al.* (2014)) or include Gaussian random effects under a multilevel representation (e.g. McNay *et al.* (2003), Dharmalingam and Morgan (2004), De Oliveira *et al.* (2014) and De Oliveira and Dias (2014)).

Although this generalization is appealing in accounting for the hierarchical structure of the data, the resulting borrowing of information and quantification of uncertainty can be quite sensitive to departures from the normality assumption (e.g. Dunson (2010)), which are arguably expected in our study. Indeed, recalling our application, it is reasonable to expect states with common unobserved characteristics such as cultural acceptance, accessibility or women's social condition to have a comparable effect on the decision process underlying the contraceptive behaviours. Moreover, such effects may vary substantially between groups of states because of the socio-economic interstate differences in India (e.g. Bala (2010)). Hence, assuming a common Gaussian distribution may force the state-specific random effects to overshrink around the population mean and rule out possible clustering among states. Consistent with this discussion, we consider a flexible representation for the prior distribution of the states-specific effects and incorporate clustering by replacing the common Gaussian assumption with a location mixture of Gaussian distributions:

$$\mu_{ik} | \Pi_k \sim \Pi_k, \quad i = 2, \ldots, 33, \qquad \Pi_k = \sum_{h=1}^{H} \nu_{hk} N(\bar{\mu}_{hk}, \sigma_k^2), \qquad \text{independently for } k = 1, \ldots, 3, \tag{6}$$

having priors for the mixing probabilities $(\nu_{1k}, \ldots, \nu_{Hk})$, and kernel parameters $\bar{\mu}_{1k}, \ldots, \bar{\mu}_{Hk}$, and σ_k^2

$$(\nu_{1k}, \ldots, \nu_{Hk}) \sim \text{Dirich}(1/H, \ldots, 1/H), \qquad \bar{\mu}_{hk} \sim N(0, \sigma_{\mu k}^2), \qquad \sigma_k^{-2} = \tau_k \sim \text{Ga}(a_{\tau_k}, b_{\tau_k}), \tag{7}$$

for every $h = 1, \ldots, H$ and $k = 1, \ldots, 3$.

A key result in prior expression (6) is that the mixture representation favours ties between state-specific effects, with states in the same cluster having the same Gaussian prior. Specifically, let G_{ik} be the cluster indicator of state i in the kth sequential logit—with $\text{pr}(G_{ik} = h) = \nu_{hk}$—the mixture of Gaussians prior favours clustering effects between the states, with $(\mu_{ik} | G_{ik} = h) \sim N(\bar{\mu}_{hk}, \sigma_k^2)$, for each $i = 2, \ldots, 33$, and $k = 1, \ldots, 3$. This property is particularly useful in our application, favouring states with common unobserved characteristics to share the same parameters. Note also that in expression (7) the mixing probabilities $(\nu_{1k}, \ldots, \nu_{Hk})$ have a Dirichlet prior with parameters $(1/H, \ldots, 1/H)$. This choice is motivated by recent theoretical results on recovering the true number of components in Gaussian mixture models (Rousseau and Mengersen, 2011). When all the mixture components—except one—are empty, representation (6) reduces to a common Gaussian prior for all the state-specific effects, so classical multilevel models for contraceptive preferences are special cases of our representation.

We seek similar flexibility also for the functional effect of the variable AGE in model (4), away from classical linear parametric representations (e.g. Husain *et al.* (2013)). In fact, as shown in Fig. 2, such an assumption may be overly restrictive in our application, thereby affecting the quality of inference. Introducing specific parameters for classes of age as in McNay *et al.* (2003), Dharmalingam and Morgan (2004), Rai and Unisa (2013), Ram *et al.* (2014), De Oliveira *et al.* (2014), De Oliveira and Dias (2014) and Mishra *et al.* (2014) improves flexibility, but questions remain on the number and location of the thresholds, which may lead to substantially different results. Motivated by this consideration, we avoid a prespecified functional form for $f_k(\cdot)$, $k = 1, \ldots, 3$, and model the unknown functions $f_k(\cdot)$ via a flexible linear combination of B-spline basis functions $\mathcal{B}_m(\cdot)$, $m = 1, \ldots, M$, obtaining

$$f_k(\text{age}_{ij}) = \sum_{m=1}^{M} \gamma_{mk} \mathcal{B}_m(\text{age}_{ij}), \qquad \text{independently for } k = 1, \ldots, 3, \tag{8}$$

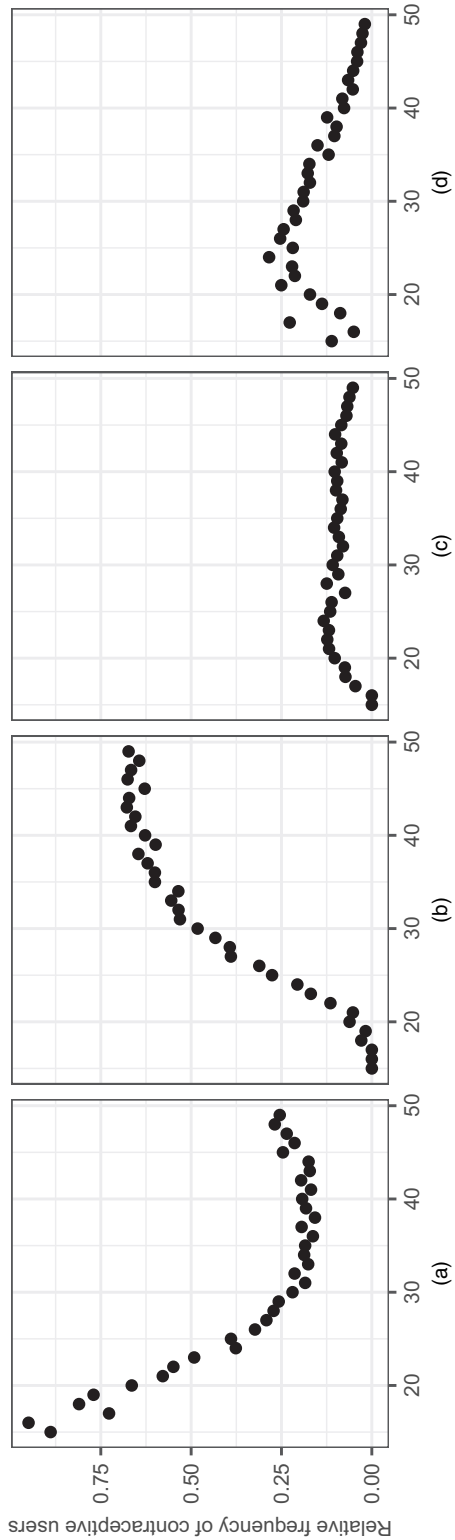

Fig. 2. Observed relative frequency of contraceptive users at different ages displayed for every contraceptive method under analysis: (a) no contraceptive, (b) sterilization (female and male), (c) natural methods (including withdrawal and periodic abstinence) and (d) modern methods (condom, oral pill, copper intrauterine device and others)

where $M = n_{\text{knots}} - 4$ is the total number of basis functions, n_{knots} denotes the total number of knots characterizing the B-spline basis and $\boldsymbol{\gamma}_k = (\gamma_{1k}, \ldots, \gamma_{Mk})^{\mathrm{T}}$ represents the vector of real coefficients governing the linear combination of the B-spline basis which characterizes $f_k(\text{age}_{ij})$.

Equation (8) defines a flexible representation for the functional effect of the variable AGE in model (4) and has been successfully considered in various demographic applications (e.g. McNeil *et al*. (1977)), beyond contraceptive studies. However, it requires the choice of the number and location of the knots, with poor choices leading to either possible bias or overfitting. To overcome this difficulty, we follow Eilers and Marx (1996) by relying on a sufficiently large number of equally spaced knots and impose the penalty $\lambda_k \Sigma_{m=3}^{M} (\gamma_{mk} - 2\gamma_{(m-1)k} + \gamma_{(m-2)k})^2$ in the log-likelihood to penalize highly parameterized representations. This penalty forces subsequent coefficients to be similar, thus controlling the smoothness of the function $f_k(\cdot)$ by an amount which depends directly on the smoothing parameter $\lambda_k > 0$. As discussed in Lang and Brezger (2004), this penalization can be rephrased within a Bayesian framework by assuming a particular rank deficient Gaussian prior for each set of coefficients $\boldsymbol{\gamma}_k$:

$$\boldsymbol{\gamma}_k | \lambda_k \sim N(0, \lambda_k^{-1} \mathbf{D}^{+}), \qquad \text{independently for } k = 1, \ldots, 3, \tag{9}$$

where $\mathbf{D}^{+}$ represents the pseudoinverse of a suitable positive semidefinite matrix $\mathbf{D}$, which is fully determined by the aforementioned quadratic penalty. Recalling Lang and Brezger (2004), such a prior distribution corresponds to a second-order Gaussian random walk, with precision λ_k. To learn the smoothness of each function, we additionally define a gamma prior $\lambda_k \sim \text{Ga}(a_\lambda, b_\lambda)$, for $k = 1, \ldots, 3$.

To conclude our Bayesian formulation we consider a multivariate Gaussian prior for the dummy coefficients $\boldsymbol{\beta}_k$, $k = 1, \ldots, 3$, in model (4) by letting

$$\boldsymbol{\beta}_k \sim N(\mathbf{b}, \mathbf{B}), \qquad \text{independently for } k = 1, \ldots, 3, \tag{10}$$

where $\mathbf{b}$ and $\mathbf{B}$ are the prior mean vector and covariance matrix respectively. Note also that, since the variables AREA, RELIGION, EDUCATION and CHILD are qualitative, characterizing these covariates via a set of dummy indicators for each category, already provides a fully flexible specification for the additive effects of AREA, RELIGION, EDUCATION and CHILD.

Beside providing a statistical model which is coherent with our research interests, the three logistic regressions that are defined in equations (5) can be studied separately—according to factorization (3). Within a Bayesian framework, this property has the key computational benefit of allowing separate Markov chain Monte Carlo algorithms for posterior computation, provided that the prior distributions (6)–(10) for the parameters in equation (4) are specified independently for $k = 1, \ldots, 3$. Although it would be possible to introduce dependence between the covariates effects also across the three logistic regressions, the increment in efficiency may be low relative to the restrictions that are induced by this higher level borrowing of information. Hence, we prefer to avoid other hierarchical layers, which may unnecessarily increase model complexity and computational intractability. In fact, as we shall discuss in Section 4, maintaining independence between the priors in the different logistic regressions does not affect efficiency and already provides an effective representation. The priors are also defined to maintain full conditional conjugacy for a simple implementation of the Gibbs sampler that is described in Section 3.

3. Posterior computation

Posterior computation relies on a recent data augmentation scheme—based on Pólya–gamma variables—which addresses the lack of conjugacy in Bayesian logistic regression (Polson *et al*.,

2013; Choi and Hobert, 2013). This approach allows simple full conditional conjugate Bayesian inference exploiting the exact representation of the binomial likelihood—parameterized via log-odds—as a scale mixture of Gaussian distributions with Pólya–gamma mixing measure. Specifically, assuming a Bayesian logistic regression with $y_i \sim \text{Bern}[\{1+\exp(-\mathbf{x}_i^{\text{T}}\boldsymbol{\beta})\}^{-1}]$, $i = 1, \ldots, n$, and $\boldsymbol{\beta} \sim N(\mathbf{b}, \mathbf{B})$, the Markov chain Monte Carlo Gibbs sampler based on the Pólya–gamma scheme exploits the fact that, given Pólya–gamma auxiliary data $\omega_i \sim \text{PG}(1, \mathbf{x}_i^{\text{T}}\boldsymbol{\beta})$, the contribution to the likelihood for the ith statistical unit is proportional to

$$\exp\left\{-\frac{\omega_i}{2}\left(\frac{y_i - \frac{1}{2}}{\omega_i} - \mathbf{x}_i^{\text{T}}\boldsymbol{\beta}\right)^2\right\}, \qquad \text{for every } i = 1, \ldots, n.$$

Hence, leveraging the above representation which relies on a more tractable Gaussian regression for transformed data $(y_i - 0.5)/\omega_i$, the resulting Gibbs sampler simply alternates between the full conditionals

$$\omega_i | \boldsymbol{\beta}, \mathbf{x}_i \sim \text{PG}(1, \mathbf{x}_i^{\text{T}}\boldsymbol{\beta})$$

and

$$\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\omega}, \mathbf{X} \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta),$$

with $\boldsymbol{\Sigma}_\beta = (\mathbf{X}^{\text{T}}\boldsymbol{\Omega}\mathbf{X} + \mathbf{B}^{-1})^{-1}$, $\boldsymbol{\mu}_\beta = \boldsymbol{\Sigma}_\beta(\mathbf{X}^{\text{T}}\boldsymbol{\eta} + \mathbf{B}^{-1}\mathbf{b})$, $\boldsymbol{\eta} = (y_1 - \frac{1}{2}, \ldots, y_n - \frac{1}{2})^{\text{T}}$ and $\boldsymbol{\Omega} = \text{diag}(\omega_1, \ldots, \omega_n)$. Efficient methods for sampling from Pólya–gamma random variables are provided in Polson *et al.* (2013) along with the reference to their implementation in the R library `BayesLogit`.

We adapt the above Gibbs sampling algorithm to incorporate functional age effects via Bayesian penalized splines, and state-specific effects whose prior distribution is a mixture of Gaussian distributions. This is accomplished by combining the Pólya–gamma data augmentation with classical Gibbs samplers for Bayesian finite mixtures of Gaussian distributions, and for Bayesian penalized splines (Lang and Brezger, 2004). Leveraging the reparameterization of the multinomial likelihood in equation (3), the Gibbs algorithms to obtain posterior samples for the parameters in representation (4) can be performed separately, and with the same derivations for each logistic regression in the sequential formulation (5).

We outline below the detailed steps to update the prior distributions for the parameters that are associated with the usage choice model. The Gibbs samplers for the reversibility and the method choice models proceed in the same way, conditioning on the appropriate statistical units. In particular, in the reversibility choice model only women using a contraceptive method are considered to update the prior distributions. Similarly, the Gibbs sampler for the parameters in the method choice model leverages only information of statistical units using contraceptives but not being sterilized.

Let $z_{ij1} = \Sigma_{r=2}^{4} y_{ijr}$ denote the binary indicator for the use of contraceptive methods, the Gibbs sampler for the parameters in the usage choice model—corresponding to $k = 1$—with priors (6)–(10), proceeds according to the following steps.

Step 1: update each Pólya–gamma augmented data ω_{ij1} from the full conditional $\omega_{ij1}|- \sim \text{PG}\{1, \mu_{i1} + f_1(\text{age}_{ij}) + \mathbf{x}_{ij}^{\text{T}}\boldsymbol{\beta}_1\}$, for every $i = 1, \ldots, 33$ and $j = 1, \ldots, n_{i1}$.
Step 2: in updating the functional effect of the variable AGE on the probability of using or not a contraception method, the full conditional distribution for the vector of parameters $\boldsymbol{\gamma}_1$ is

$$\boldsymbol{\gamma}_1|- \sim N\{(\mathbf{H}_1^{\text{T}}\boldsymbol{\Omega}_1\mathbf{H}_1 + \lambda_1\mathbf{D})^{-1}\mathbf{H}_1^{\text{T}}\boldsymbol{\eta}_{\gamma_1}, (\mathbf{H}_1^{\text{T}}\boldsymbol{\Omega}_1\mathbf{H}_1 + \lambda_1\mathbf{D})^{-1}\},$$

where $\boldsymbol{\eta}_{\gamma_1}$ is a vector with entries $\eta_{ij\gamma_1} = z_{ij1} - \frac{1}{2} - \omega_{ij1}(\mu_{i1} + \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1)$, $i = 1, \ldots, 33$, $j = 1, \ldots, n_{i1}$, whereas $\mathbf{H}_1$ is the B-splines design matrix having row entries $\mathcal{B}_m(\mathrm{age}_{ij})$, $i = 1, \ldots, 33$, $j = 1, \ldots, n_{i1}$, for every column $m = 1, \ldots, M$. Finally, $\boldsymbol{\Omega}_1$ is a diagonal matrix with entries ω_{ij1}, $i = 1, \ldots, 33$, $j = 1, \ldots, n_{i1}$, on its diagonal.

Step 3: update the parameter λ_1, controlling the smoothness for the functional effect of the variable AGE, from its full conditional gamma random variable $\lambda_1 | - \sim \mathrm{Ga}\{a_\lambda + \mathrm{rank}(\mathbf{D})/2, b_\lambda + \boldsymbol{\gamma}_1^{\mathrm{T}}\mathbf{D}\boldsymbol{\gamma}_1/2\}$. Note that $\mathbf{D}$ is not a full rank matrix, and therefore $\mathrm{rank}(\mathbf{D}) = M - 2$.

Step 4: exploiting the Pólya–gamma data augmentation scheme, the full conditional distribution for the vector of parameters $\boldsymbol{\beta}_1$ encoding the effect of the dummy covariates on the probability of using or not a contraceptive is

$$\boldsymbol{\beta}_1 | - \sim N\{(\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\Omega}_1\mathbf{X}_1 + \mathbf{B}^{-1})^{-1}(\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\eta}_{\beta_1} + \mathbf{B}^{-1}\mathbf{b}), (\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\Omega}_1\mathbf{X}_1 + \mathbf{B}^{-1})^{-1}\},$$

where $\boldsymbol{\eta}_{\beta_1}$ is a vector with entries $\eta_{ij\beta_1} = z_{ij1} - \frac{1}{2} - \omega_{ij1}\{\mu_{i1} + f_1(\mathrm{age}_{ij})\}$, $i = 1, \ldots, 33$, $j = 1, \ldots, n_{i1}$, whereas $\mathbf{X}_1$ is the corresponding design matrix having row entries $\mathbf{x}_{ij}^{\mathrm{T}}$.

Step 5: to update the state-specific parameters under prior (6)–(7), we combine the Pólya–gamma data augmentation with classical algorithms for Bayesian mixtures of Gaussian distributions. In particular—leveraging the mixture representation in equation (6)—we first allocate each state $i = 2, \ldots, 33$, to one of the $h = 1, \ldots, H$ clusters by sampling each group indicator G_{i1}, $i = 2, \ldots, 33$, from the full conditional categorical random variable with probabilities

$$\mathrm{pr}(G_{i1} = h | -) = \frac{\nu_{h1} N(\mu_{i1}; \bar{\mu}_{h1}, \sigma_1^2)}{\sum_{s=1}^{H} \nu_{s1} N(\mu_{i1}; \bar{\mu}_{s1}, \sigma_1^2)}, \qquad h = 1, \ldots, H.$$

Step 6: consistent with prior distribution (6), $(\mu_{i1} | G_{i1} = h) \sim N(\bar{\mu}_{h1}, \sigma_1^2)$. Hence, the full conditional of each state-specific parameter μ_{i1}, given the above cluster assignments, is easily available as

$$\mu_{i1} | - \sim N\left(\frac{\sum_{j=1}^{n_{i1}} [z_{ij1} - \frac{1}{2} - \omega_{ij1}\{f_1(\mathrm{age}_{ij}) + \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1\}] + \bar{\mu}_{G_{i1}1}/\sigma_1^2}{1/\sigma_1^2 + \sum_{j=1}^{n_{i1}} \omega_{ij1}}, \frac{1}{1/\sigma_1^2 + \sum_{j=1}^{n_{i1}} \omega_{ij1}} \right),$$

independently for every $i = 2, \ldots, 33$.

Step 7: since σ_1^2 is shared between all the mixture components, and provided that $(\mu_{i1} | G_{i1} = h) \sim N(\bar{\mu}_{h1}, \sigma_1^2)$, the full conditional for $\sigma_1^{-2} = \tau_1$ is $\tau_1 | - \sim \mathrm{Ga}\{a_{\tau_1} + 32/2, b_{\tau_1} + \Sigma_{i=2}^{33}(\mu_{i1} - \bar{\mu}_{G_{i1}1})^2/2\}$.

Step 8: exploiting again the result $(\mu_{i1} | G_{i1} = h) \sim N(\bar{\mu}_{h1}, \sigma_1^2)$, the full conditional for $\bar{\mu}_{h1}$ is

$$\bar{\mu}_{h1} | - \sim N\left(\frac{\sum_{i:G_{i1}=h} \mu_{i1}/\sigma_1^2}{1/\sigma_{\mu 1}^2 + \sum_{i:G_{i1}=h} 1/\sigma_1^2}, \frac{1}{1/\sigma_{\mu 1}^2 + \sum_{i:G_{i1}=h} 1/\sigma_1^2} \right),$$

independently for every $h = 1, \ldots, H$.

Step 9: update the mixing probability vector $(\nu_{11}, \ldots, \nu_{H1})$ from its full conditional Dirichlet distribution $\mathrm{Dirich}\{1/H + \Sigma_{i=2}^{33}\mathbf{1}(G_{i1} = 1), \ldots, 1/H + \Sigma_{i=2}^{33}\mathbf{1}(G_{i1} = H)\}$.

4. Sociodemographic determinants underlying the contraceptive choices in India

Recalling our research interests, we apply the sequential logistic regressions that are outlined in representation (5) to the contraceptive preference data that were described in Section 1.1. Our main goal is to estimate flexibly and to interpret the covariates effects on the binary choices characterizing the sequential decision process in Fig. 1, and to quantify the uncertainty of our conclusions. Before discussing our findings in Section 4.2, we first compare the model that was proposed in Section 2 with alternative parametric specifications to assess whether the Bayesian semiparametric model that was outlined in Section 2.2 effectively improves inference in the motivating application that is considered. We also study performance in out-of-sample prediction of contraceptive preferences and compare the results with state of the art methods specifically developed for classification tasks, to assess to what extent our Bayesian semiparametric model induces a flexible representation of the sociodemographic factors underlying the contraceptive choices. These assessments are described in detail in Section 4.1.

In performing Bayesian inference we rely on the priors (6)–(10), described in Section 2.2. As there are at most 32 clusters among the 32 parameters $\mu_{2k}, \ldots, \mu_{33k}$, characterizing the state-specific effects in each logistic regression, we fix $H = 32$ and allow the sparse Dirichlet prior (7) to delete redundant mixture components (Rousseau and Mengersen, 2011). In setting $\sigma^2_{\mu k}$ and (a_{τ_k}, b_{τ_k}) in prior (7), note that $\sigma^2_{\mu k}$ measures the variability of the component-specific mean parameters with respect to 0, whereas (a_{τ_k}, b_{τ_k}) controls the variance σ^2_k of the state-specific effects in each mixture component. For instance, high values of $\sigma^2_{\mu k}$ combined with a small σ^2_k characterize situations in which states within the same cluster have highly similar state-specific effects, but these effects substantially change between different clusters. Hence, in setting such hyperparameters, we consider a data-driven approach. Specifically, we first estimate a classical logistic regression with state-specific fixed effects and then cluster these estimated effects via standard agglomerative methods. Based on the empirical clusters, (a_{τ_k}, b_{τ_k}) are set to centre the prior for σ^2_k near the average of the within-cluster sample variance, whereas $\sigma^2_{\mu k}$ is matched with the average of the squared deviations of the cluster means from 0. Consistent with this procedure, we set $\sigma^{-2}_{\mu 1} = 0.2$, $\sigma^{-2}_{\mu 2} = 0.02$, $\sigma^{-2}_{\mu 3} = 0.01$ and $(a_{\tau_1} = 0.5, b_{\tau_1} = 0.1)$, $(a_{\tau_2} = 0.1, b_{\tau_2} = 0.1)$ and $(a_{\tau_3} = 0.15, b_{\tau_3} = 0.1)$. Although other settings are possible—using for example prior knowledge of interstate differences in India—such an empirical Bayes approach is typically useful in Bayesian hierarchical models to improve mixing and convergence. We also attempted inference under moderate changes of the above settings avoiding data-driven priors but found no evident differences in the final results thanks to the borrowing of information that is induced by the mixture of Gaussian distributions.

In defining the functional effect of the variable AGE in representation (8), we follow instead Eilers and Marx (1996) by relying on a sufficiently large number $n_{\text{knots}} = 46$ of equally spaced knots and facilitate moderate shrinkage by letting $a_\lambda = 1.5$ and $b_\lambda = 5 \times 10^{-4}$—consistent with Lang and Brezger (2004) and the results in Fig. 2. Finally, although the prior mean vector $\mathbf{b}$, and the covariance matrix $\mathbf{B}$, could be set according to current knowledge of the effects of the qualitative covariates, we let $\mathbf{b} = \mathbf{0}$ and $\mathbf{B} = \text{diag}(100, \ldots, 100)$, to incorporate the neutral hypothesis of no relevant effects, with a moderate prior uncertainty, since there is not overall agreement in current studies on $\boldsymbol{\beta}_k$. Also in this case we found posterior inference robust to moderate changes in $\mathbf{b}$ and $\mathbf{B}$. This is because the variables AREA, RELIGION, EDUCATION and CHILD have a small number of well-represented categories. Hence, there is sufficient information in the data to provide robust inference on $\boldsymbol{\beta}_k$.

In performing posterior inference we consider 22000 Gibbs iterations, holding out the first 2000 as burn-in, and thinning the chains every five samples. Consistent with standard diagnostics of Markov chain Monte Carlo algorithms—including our Gibbs sampler—we check convergence via graphical analysis of the trace plots for the posterior samples of the parameters and study mixing via effective sample sizes (Plummer *et al.*, 2006). Such diagnostics highlighted good mixing and no evidence against convergence.

4.1. Model comparisons and out-of-sample predictive performance

As discussed in Section 2.2, an important contribution of the statistical model proposed is in improving flexibility compared with current analyses of contraceptive preferences in India. This increased flexibility is motivated by the data under analysis and is accomplished by considering mixtures of Gaussian distributions for the state-specific effects, along with penalized splines for the functional effects of the variable AGE.

4.1.1. Comparison with submodels via information criteria

To assess empirically the practical usefulness of the formulation proposed, we compare our mixture splines model with simpler submodels, using the deviance information criterion DIC and Watanabe–Akaike information criterion WAIC (Gelman *et al.*, 2014). Consistent with the above considerations, and with the discussions in Section 2.2, we consider three submodels. These three alternative specifications comprise a baseline model in which the variable AGE enters the predictor linearly, and the parameters $\mu_{2k}, \ldots, \mu_{33k}$ have classical Gaussian priors; a splines model replacing the linearity assumption in the baseline model with the spline representation (8) and a mixture model in which mixtures of Gaussian distributions (6)–(7) are considered for $\mu_{2k}, \ldots, \mu_{33k}$, but the variable AGE enters the predictor linearly as in the baseline model.

The above simpler submodels are special and more parsimonious versions of our semiparametric specification, thereby proving relevant alternative representations to assess the actual usefulness of the increased flexibility that is provided by the mixture–splines model. Posterior inference for these submodels proceeds under minor modifications of the Gibbs sampler that was proposed in Section 3, and the hyperparameters are set according to the same guidelines as considered for our mixture–splines model. Note also that the simple factorization (3) of the full model likelihood, together with the independence of the priors for $k = 1, \ldots, 3$, enables us to evaluate the partial DIC and WAIC for each logistic regression in equations (5), and then to obtain those for the full model by simple summation of the partial models.

As shown in Table 1, DIC and WAIC are on a similar scale. More evident improvements are obtained when modelling the functional effect of the variable AGE under a spline representation as in equation (8), instead of a linear representation. This supports our choice of improving flexibility in characterizing non-linear effects of the variable AGE. When comparing the splines model with our mixture–splines representation, we observe also an advantage in using mixtures of Gaussian distributions instead of Gaussian priors for the state-specific effects, thereby confirming the usefulness of our semiparametric specification. This additional improvement is less evident compared with the introduction of non-linear effects for the variable AGE, meaning that a common Gaussian prior is reasonable for several states, but a subset of them may still present notable deviations from this assumption. Hence, incorporating this behaviour in our model can provide key insights on relevant deviations for groups of states from shared structures.

4.1.2. Out-of-sample predictive performance

The results in Table 1 confirm that our Bayesian semiparametric formulation is empirically preferred over simpler specifications but do not guarantee that the model proposed provides an accurate representation of the sociodemographic factors underlying the contraceptive preferences. For instance, although we improve flexibility via Bayesian splines and mixture modelling, the additive assumption for the effects of the different variables in equations (5) may still provide a restrictive representation of the determinants driving the contraceptive choices.

To understand whether the statistical model proposed is sufficiently flexible, we study the performance of our representation in out-of-sample prediction of the contraceptive preferences and compare the results with those obtained under benchmark methods for classification—e.g. discriminant analysis, random forests and gradient boosting—using the same covariates. These methods have been specifically developed to enable accurate predictions of a response variable leveraging much complex partitions of the covariates space. Therefore, a similar predictive performance under our model would provide relevant insights into the sufficient flexibility of the representation proposed, and its adequacy in accurately characterizing the sociodemographic factors underlying the contraceptive preferences.

As shown in Table 2, the predictive checks proceed with reference to two nested partitions of the response variable. The reason is that we additionally aim to compare predictive performance with a formulation recalling that proposed by De Oliveira *et al*. (2014), who focused on comparing natural and modern methods against sterilization, only for women currently using contraceptives. Consistent with this, we first study the predictive performance that is associated with the usage choice model alone—thereby focusing on the probability that a new individual will use contraceptive methods or not. Then, in the subsequent assessment, we study performance in predicting use of natural methods, modern methods or sterilization for the subset of women using contraceptives, consistent with De Oliveira *et al*. (2014). Both assessments are made on a subset of randomly selected women—comprising 25% of the sample—using the remaining statistical units as a training set for the various methods.

Out-of-sample predictions under our model formally rely on the expectation of the posterior predictive distribution of the contraceptive preference indicators, which coincides with the posterior mean of the conditional probabilities of interest. In particular, in the first assessment, we compute for every out-of-sample unit the posterior mean of the associated probability of using, ρ_{ij1}, or not using, $1-\rho_{ij1}$, a contraceptive method, and then predict the final outcome by checking whether this estimated probability exceeds or not a specific cut-off. A similar procedure holds for the second assessment, except for focusing on the posterior mean of $1-\rho_{ij2}$, $\rho_{ij2}\rho_{ij3}$ and $\rho_{ij2}\rho_{ij4}$, measuring the conditional probabilities of sterilization, modern and natural meth-

Table 1. DIC and WAIC for competing submodels†

Criterion	*Results for the following submodels:*			
	Baseline	*Splines*	*Mixture*	*Mixture–splines*
DIC	53507.70	53092.90	53505.00	*53091.25*
−2 WAIC	53503.41	53088.62	53498.06	*53083.61*

†The value −2 WAIC is reported to obtain indices on the same scale.
Values in italics are the lowest (best) DIC and −2 WAIC.

Table 2. For our model and relevant competitors, out-of-sample performance in predicting the use or not use of contraceptive methods—measured via AUC (the area under the receiver operating characteristic curve) and the misclassification rate with cut-off 0.5, and out-of-sample performance in predicting use of natural methods or modern methods or sterilization, for those women using contraceptives

	Results for the following methods:				
	Mixture–splines	*Gradient boosting*	*Random forest*	*Linear discriminant analysis*	*Multinomial*
Predictive performance for usage choice					
AUC	0.799	0.803	0.797	0.790	—
Misclassification rate	0.197	0.194	0.196	0.196	—
Combined predictive performance for reversibility and method choice					
Misclassification rate	0.230	0.231	0.234	0.249	0.233

ods respectively, given the decision to use contraceptives. In this case, the predicted value is the category having the highest estimated probability.

As shown in Table 2, although our statistical model is mainly focused on providing interpretable inference for the determinants underlying contraceptive preferences in India, we obtain a predictive performance in line with the benchmark methods that were specifically developed for prediction tasks. Moreover, the misclassification rate of our Bayesian semiparametric formulation is slightly lower than existing parametric multinomial models, such as that proposed by De Oliveira *et al*. (2014), thus suggesting that more flexible specifications are indeed preferred. In fact, in implementing a similar version of the multinomial regression in De Oliveira *et al*. (2014) we consider a piecewise constant specification for AGE in the intervals [15, 25], (25, 34] and (34, 49], which assumes that the effect of the variable AGE is the same within each interval.

Although the above assessments are based on a simple hold-out approach, it is worth noting that the accurate predictive performance is a side benefit of our statistical model, and the overarching focus is on providing meaningful and accurate inference. Hence, consistent with our fundamental goal, we avoid further complications via cross-validation and leverage the results in Table 2 to obtain a simple reassurance that the Bayesian semiparametric representation proposed does not lead to inadequate characterization of the sociodemographic factors underlying the contraceptive preferences. Moreover, we obtained similar conclusions when considering different training and test sets.

4.2. Interpretation of the results

The results in Section 4.1 confirm that the Bayesian semiparametric model that was proposed in Section 2.2 provides a sufficiently flexible and empirically motivated representation for the determinants underlying the sequential decision process that was discussed in Section 2.1, with a specific reference to the current family planning programmes in India. These results motivate discussion and interpretation of our findings via inference on the posterior distributions for the parameters in equations (5).

The posterior distributions for the effects of variables AREA, RELIGION, EDUCATION and CHILD are summarized in Table 3 and provide interesting findings on the determinants

Table 3. Posterior mean and 0.95 credible intervals for the β_k-parameters in the sequential logistic regressions†

Variable	*Results for the following choices:*		
	Usage	*Reversibility*	*Method*
Variable AREA; reference category rural			
urban	*0.24* [0.17; 0.31]	*0.29* [0.21; 0.38]	*0.45* [0.32; 0.58]
Variable RELIGION; reference category hindu			
muslim	*−0.43* [−0.52; −0.34]	*1.24* [1.12; 1.36]	0.13 [−0.03; 0.29]
christian	*−0.26* [−0.48; −0.03]	0.00 [−0.29; 0.29]	0.36 [−0.13; 0.88]
other	0.08 [−0.11; 0.29]	*0.46* [0.25; 0.66]	*0.30* [0.02; 0.59]
Variable EDUCATION; reference category no education			
low	*0.14* [0.05; 0.23]	0.08 [−0.03; 0.19]	*0.43* [0.25; 0.60]
intermediate	*0.20* [0.12; 0.27]	*0.50* [0.40; 0.59]	*0.71* [0.56; 0.86]
high	*0.27* [0.16; 0.37]	*1.28* [1.14; 1.41]	*1.16* [0.97; 1.34]
Variable CHILD; reference category more than one child			
no child	*−3.71* [−3.88; −3.54]	*2.20* [1.69; 2.75]	−0.19 [−0.61; 0.22]
one child	*−1.37* [−1.45; −1.28]	*2.19* [2.06; 2.32]	*−0.19* [−0.34; −0.04]

†The parameter estimates in italics have 0.95 credible intervals not including the value 0.

of the contraceptive preferences. Lack of information about family planning, inaccessibility issues, a marked preference for sons (e.g. Chavada and Bhagyalaxmi (2009)) and decreased woman empowerment (e.g. Lee-Rife (2010)) in rural areas motivate a lower probability of contraceptive use—compared with urban areas—along with a reduced preference for reversible methods instead of sterilization. Consistent with this result, the preference for modern temporary methods—compared with natural methods—increases in urban areas with respect to rural areas. This is also in line with reduced knowledge in rural areas about condoms and their additional effect in preventing sexually transmitted diseases (e.g. Donta *et al*. (2014)).

Focusing on the religion effect, there is a literature providing comparisons between Muslims and Hindus with respect to contraceptive behaviour (e.g. Dharmalingam and Morgan (2004)). The substantially different fertility intentions between these two religions motivates a reduced use of contraceptives for Muslims—compared with Hindus. Additionally, Islamic opposition to sterilization is evident in the reversibility choice with an increased preference for temporary contraceptives compared with Hindus—among partners using contraceptive methods. Christians are instead more similar to Hindus, with the exception of a reduced attitude towards the use of contraceptives. This result is in line with the lack of formal prohibitions with respect to contraceptions in Hinduism, making this practice acceptable. Refer also to Srikanthan and Reid (2008) for an additional discussion of religion influences on contraception.

Consistent with recent contributions (e.g. Rizwan *et al*. (2012)), growing literacy has an increasing positive effect in the decision to use contraceptives and in avoiding sterilization—among partners considering contraceptive methods. These effects are reasonably favoured by higher accessibility, better knowledge of contraception, lower preference for sons and an increasing possibility of women empowerment within the household. Highly educated individuals are further characterized by an increased preference for modern methods—among partners opting for temporary contraceptives. This finding is in line with an increased awareness on sexually transmitted diseases (e.g. Donta *et al*. (2014)).

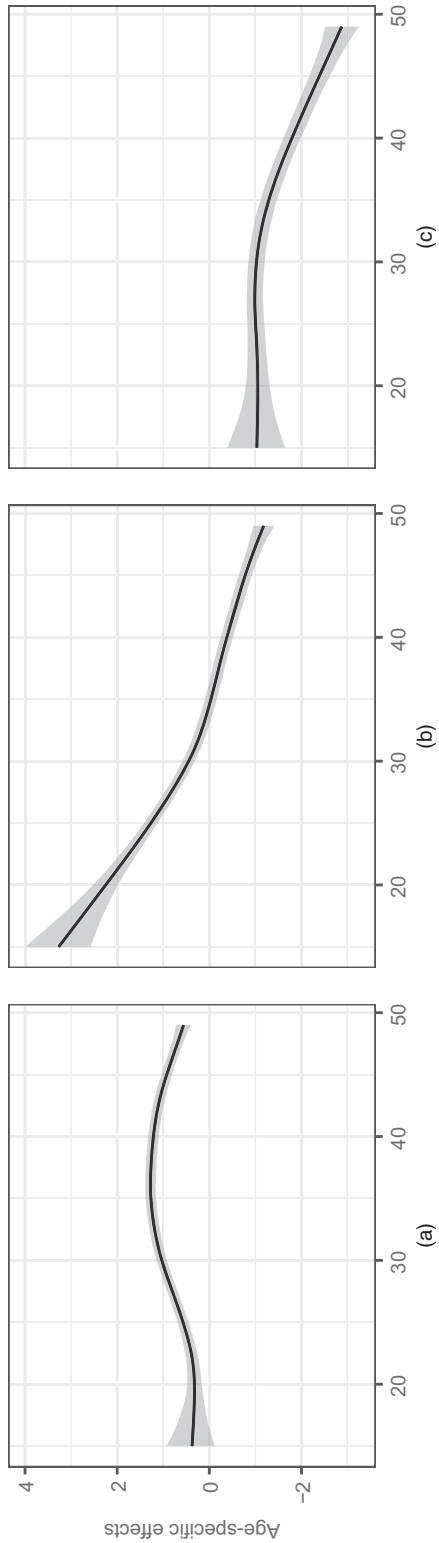

Fig. 3. Posterior mean and pointwise 0.95 credible intervals (■) for the functional effect of the variable AGE in the sequential logistic regressions for (a) usage choice, (b) reversibility choice and (c) method choice

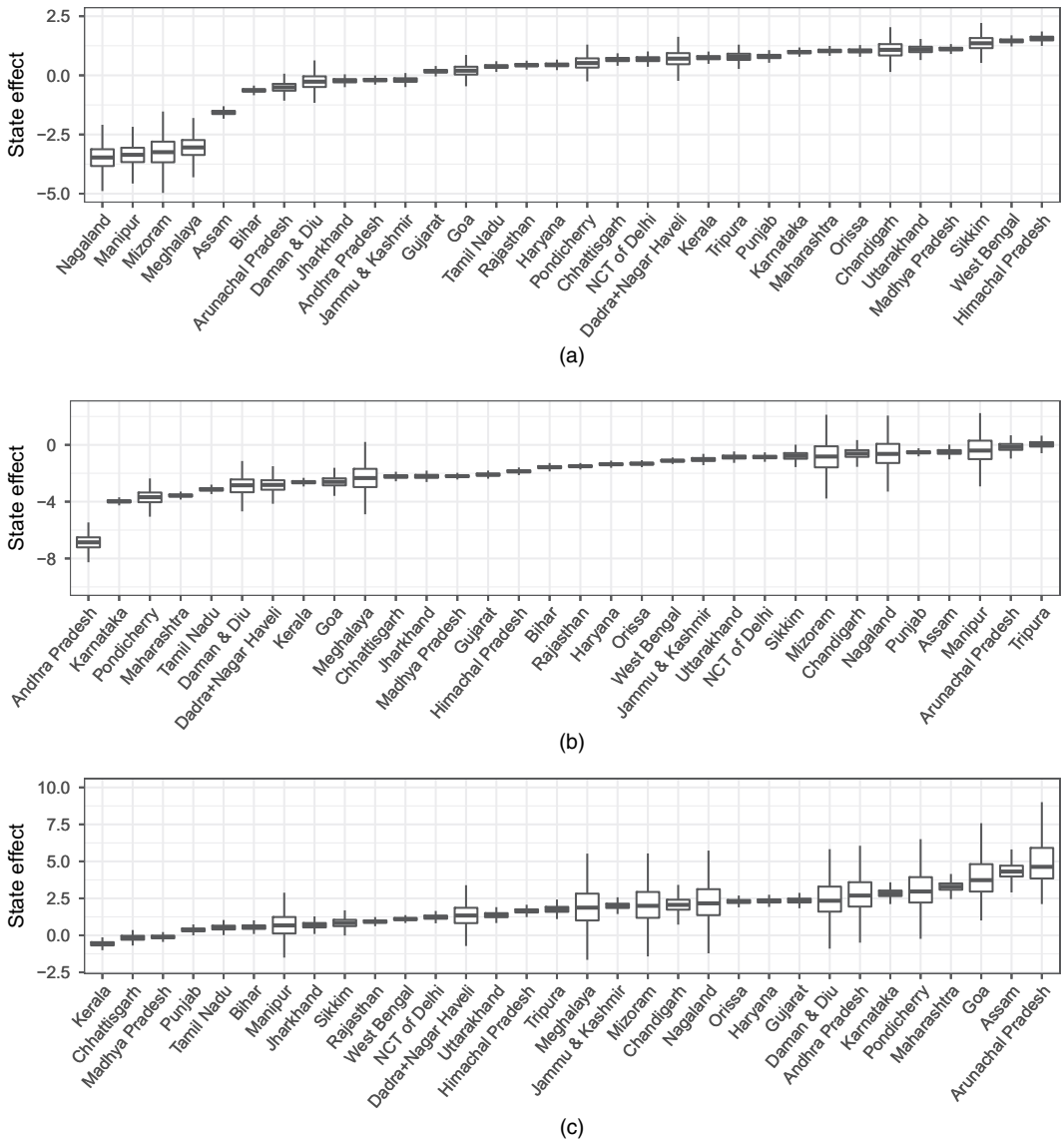

Fig. 4. Boxplots summarizing the posterior distribution of the state-specific effects in the sequential logits for (a) usage choice, (b) reversibility choice and (c) method choice

Finally—compared with women having more than one child—we observe an increasingly lower preference towards contraceptive use and sterilization in the sequential process characterizing women with no or one child. These results are in line with recent findings and are associated with the different fertility intentions and son preferences at varying number of children (e.g. Das (1986)).

Fig. 3 summarizes the posterior distribution for the functional effect of the variable AGE at the various steps of the decision process underlying contraceptive preferences. As expected, the effect on the use of contraceptives has an overall parabolic trend, peaking between 30 and 40 years when birth control is more common. Sterilization is still the most common contraceptive

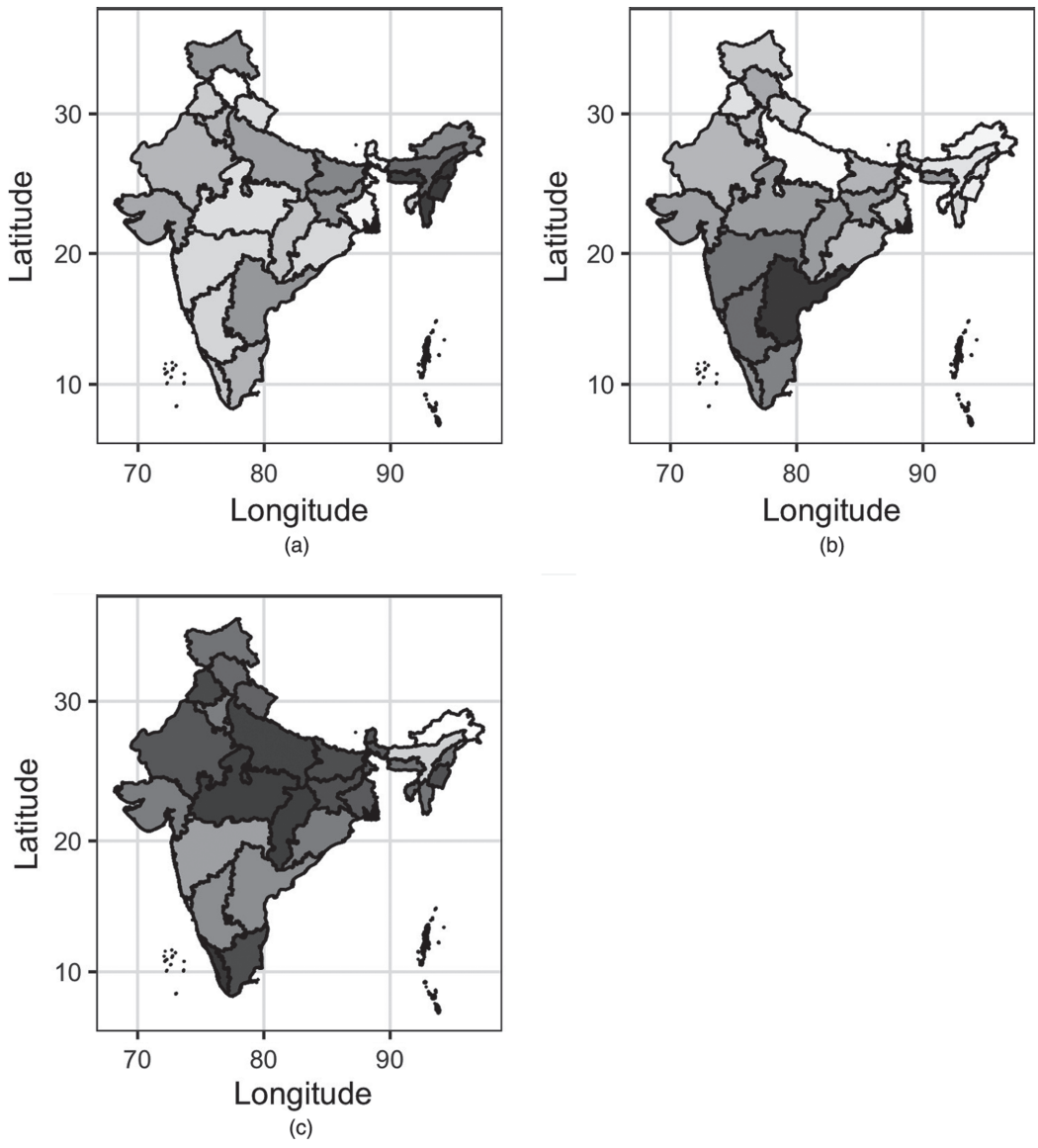

Fig. 5. States of India coloured according to their corresponding estimated effect in each logistic regression for the various steps of the sequential decision process in Fig. 1: (a) usage choice, (b) reversibility choice and (c) method choice

method in India (e.g. Pachauri (2004, 2014)) and its prevalence tends to increase with age (e.g. Säävälä (1999)), motivating the previously discussed trend along with the decreasing functional effect of the variable AGE on the choice of temporary methods instead of sterilization. Among the women opting for reversible methods, we observe an increasing preference towards natural strategies—compared with modern methods—as age grows. This trend is evident only after 30 years, meaning that young couples have still, potentially unmet, interest in modern methods (e.g. Pachauri (2004)).

These results are mostly in line with recent studies on contraceptive preferences in India (e.g. McNay *et al*. (2003), Husain *et al*. (2013), Rai and Unisa (2013), De Oliveira *et al*. (2014), De Oliveira and Dias (2014), Ram *et al*. (2014), Mishra *et al*. (2014) and Haque and Patel (2015)). However, as already discussed in Sections 1 and 2, our model provides a more global and flexible overview which is motivated by current family planning policies in India and avoids focusing only on specific aspects of contraceptive preferences in India.

To conclude our study, we focus on the posterior distributions of the state-specific incremental effects with respect to the baseline Uttar Pradesh. Socio-economic differences across states—or groups of states—are marked in India (e.g. Bala (2010) and Das (1999)). Current studies exploring state-specific differences in contraceptive preferences typically focus on subsamples of states (e.g. Rai and Unisa (2013)), or groups of states (e.g. De Oliveira *et al*. (2014) and Ram *et al*. (2014)) or estimate different models for each state (e.g. Dharmalingam and Morgan (2004)). Dharmalingam and Morgan (2004) considered also a multilevel analysis, but their focus was on religious differences in contraceptive preferences. Our model allows instead inference also on state-specific effects after controlling for other covariates.

The benefits that are associated with our mixture of Gaussians prior (6)–(7) for the state-specific parameters are clear in Figs 4 and 5, which show groups of states whose effects are not forced to overshrink around the global mean and display clustering effects that are interestingly in line with the geographical positions of the different states—without informing the model of such geographical structure. This improved flexibility highlights a substantially lower intention towards contraceptive use for a group of north-eastern states including Nagaland, Manipur, Mizoram and Meghalaya. This clustering tendency for the north-eastern states is also evident in the logistic regression that is associated with the choice of reversibility and is further confirmed when applying the procedure of Medvedovic and Sivaganesan (2002) for clustering in mixture models. This result is in line with the common political history and specific cultural aspects of the north-eastern states in India, which are also referred to as 'seven sisters' (Baruah, 2006). Andhra Pradesh displays instead a substantially lower preference for temporary methods, compared with sterilization. This confirms the strong measures that were adopted by the government of Andhra Pradesh to promote sterilization (Prakasamma, 2009).

Finally, it is also worth noting that in Fig. 4 some boxplots are wider than others, because some states have reduced information for specific contraceptive preferences. This result further motivates our choice of improving borrowing of information via a Bayesian mixture of Gaussian distributions for the state-specific effects, which still maintains flexibility in modelling more evident deviations.

5. Conclusion

In India contraceptive preferences are subject to a complex combination of family planning policies and sociodemographic differences, thereby requiring meaningful statistical models and flexible inference procedures to provide interpretable and accurate conclusions. The available statistical models are not sufficiently flexible and typically fail to provide a global overview of the determinants underlying the entire decision process characterizing the choices of contraceptive. To address this gap, we developed a Bayesian semiparametric statistical model relying on a set of logistic regressions which characterize a sequential decision process motivated by the current family planning policies in India.

Our results substantially agree with the descriptive analyses that are available from other national surveys such as the National Family Health Survey and are typically in line with findings from other contributions studying only a subset of the entire decision process in Fig. 1.

A major benefit of our formulation is in allowing inference and quantification of uncertainty on the entire set of contraceptive preferences, disambiguating the effects of the sociodemographic covariates at every step of the sequential process in Fig. 1, within a unique and flexible statistical model. This approach to inference provides a global and interpretable overview of the entire contraceptive preferences, facilitating the assessment of current family planning policies, and an improved targeting of subpopulations not yet addressed.

Although other decision mechanisms could be considered, the process in Fig. 1 is of interest as discussed in Section 2.1 and provides a formal reparameterization of the multinomial probability mass function for contraceptive preferences. This facilitates also posterior inference on the contraceptive probabilities in equation (1) for each configuration of covariates characterizing different women's profiles.

Acknowledgements

We thank the Joint Editor, the Associate Editor and the three referees for their valuable comments. This work was partially funded by grant CPDA154381/15 from the University of Padova, Italy. The original collector of the data, the Inter-University Consortium for Political and Social Research, and the relevant funding agency bear no responsibility for use of the data or for interpretations or inferences based on such uses.

References

Agresti, A. (2007) *An Introduction to Categorical Data Analysis*. New York: Wiley.

Alkema, L., Kantorova, V., Menozzi, C. and Biddlecom, A. (2013) National, regional, and global rates and trends in contraceptive prevalence and unmet need for family planning between 1990 and 2015: a systematic and comprehensive analysis. *Lancet*, **381**, 1642–1652.

Bala, N. (2010) Inter-state imbalances in social and economic development of India. *J. Res. Natn. Devlpmnt*, **8**.

Baruah, S. (2005) *Durable Disorder: Understanding the Politics of Northeast India*. Oxford: Oxford University Press.

Bijak, J. and Bryant, J. (2016) Bayesian demography 250 years after Bayes. *Popln Stud.*, **70**, 1–19.

Bloom, D. E. (2012) Population dynamics in India and implications for economic growth. In *The Oxford Handbook of the Indian Economy* (ed. C. Ghate), pp. 462–498. Oxford: Oxford University Press.

Chaurasia, A. R. and Singh, R. (2014) Forty years of planned family planning efforts in India. *J. Famly Welfr.*, **60**, 1–16.

Chavada, M. and Bhagyalaxmi, A. (2009) Effect of socio-cultural factors on the preference for the sex of children by women in Ahmedabad district. *Hlth Popln Perspect. Iss.*, **32**, 184–189.

Choi, M. H. and Hobert, J. P. (2013) The Pólya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electron. J. Statist.*, **7**, 2150–2163.

Das, A. (1999) Socio-economic development in India: a regional analysis. *Devlpmnt Soc.*, **28**, 313–345.

Das, E. (1986) The sex of previous children and subsequent fertility intention in India. *Can. Stud. Popln*, **13**, 19–36.

De Oliveira, I. T. and Dias, J. G. (2014) Disentangling the relation between wealth and contraceptive use in India: a multilevel probit regression approach. *Qual. Quant.*, **48**, 1001–1012.

De Oliveira, I. T., Dias, J. G. and Padmadas, S. S. (2014) Dominance of sterilization and alternative choices of contraception in India: an appraisal of the socioeconomic impact. *PLOS One*, **9**, no. 6, article e100050.

Desai, S., Vanneman, R. and National Council of Applied Economic Research (2015) India Human Development Survey-II (IHDS-II), 2011-12. Inter-University Consortium for Political and Social Research, Ann Arbor.

Dharmalingam, A. and Morgan, S. P. (2004) Pervasive muslim-hindu fertility differences in India. *Demography*, **41**, 529–545.

Donta, B., Begum, S. and Naik, D. D. (2014) Acceptability of male condom: an Indian scenario. *Ind. J. Med. Res.*, **120**, 152–156.

Dunson, D. B. (2010) Nonparametric Bayes applications to biostatistics. In *Bayesian Nonparametrics* (eds N. L. Hjort, C. Holmes, P. Müller and S. G. Walker), pp. 223–273. New York: Cambridge University Press.

Eilers, P. and Marx, B. (1996) Flexible smoothing using B-splines and penalized likelihood. *Statist. Sci.*, **11**, 89–121.

Elderd, B. D. and Miller, T. E. (2016) Quantifying demographic uncertainty: Bayesian methods for integral projection models. *Ecol. Monogr.*, **86**, 125–144.

Gelman, A., Hwang, J. and Vehtari, A. (2014) Understanding predictive information criteria for Bayesian models. *Statist. Comput.*, **24**, 997–1016.
Haque, I. and Patel, P. P. (2015) Socioeconomic and cultural differentials of contraceptive usage in West Bengal: evidence from National Family Health Survey Data. *J. Famly Hist.*, **40**, 230–249.
Harkavy, O. and Roy, K. (2007) The emergence of the Indian National Family Planning Program. In *The Global Family Planning Revolution, Three Decades and Policies and Programs* (eds W. C. Robinson and J. A. Ross), pp. 301–323. Washington DC: World Bank.
Husain, Z., Ghosh, S. and Dutta, M. (2013) 'Ultramodern contraception' re-examined: cultural dissent or son preference? *Asn Popln Stud.*, **9**, 280–300.
Jayaraman, A., Mishra, V. and Arnold, F. (2009) The relationship of family size and composition to fertility desires, contraceptive adoption and method choice in South Asia. *Int. Perspect. Sexl Reproduct. Hlth*, **35**, 29–38.
Lang, S. and Brezger, B. (2004) Bayesian P-splines. *J. Computnl Graph. Statist.*, **13**, 183–212.
Lee-Rife, S. M. (2010) Women empowerment and reproductive experiences over the life course. *Socl Sci. Med.*, **71**, 634–642.
Li, Y., Müller, P. and Lin, X. (2011) Center-adjusted inference for a nonparametric Bayesian random effect distribution. *Statist. Sin.*, **21**, 1201–1223.
Mauldin, W. P. and Segal, S. J. (1988) Prevalence of contraceptive use: trends and issues. *Stud. Famly Planng*, **19**, 335–353.
McNay, K., Arokiasamy, P. and Cassen, R. (2003) Why are uneducated women in India using contraception?: A multilevel analysis. *Popln Stud.*, **57**, 21–40.
McNeil, D. R., Trullell, T. J. and Turner, J. C. (1977) Spline interpolation of demographic data. *Demography*, **14**, 245–252.
Medvedovic, M. and Sivaganesan, S. (2002) Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, **18**, 1194–1206.
Mishra, A., Nanda, P., Speizer, I. S., Calhoun, L. M., Zimmerman, A. and Bhardwaj, R. (2014) Mens attitudes on gender equality and their contraceptive use in Uttar Pradesh India. *Reprod. Hlth*, **11**, article 41.
Pachauri, S. (2004) Expanding contraceptive choice in India: issues and evidence. *J. Famly Welfr*, **50**, 13–25.
Pachauri, S. (2014) Priority strategies for India's family planning programme. *Ind. J. Med. Res.*, **140**, 137–146.
Plummer, M., Best, N., Cowles, K. and Vines, K. (2006) CODA: convergence diagnosis and output analysis for MCMC. *R News*, **6**, 7–11.
Polson, N. G., Scott, J. G. and Windle, J. (2013) Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Am. Statist. Ass.*, **108**, 1339–1349.
Prakasamma, M. (2009) Maternal mortality-reduction programme in Andhra Pradesh. *J. Hlth Popln Nutrn*, **27**, 220–234.
Rai, R. K. and Unisa, S. (2013) Dynamics of contraceptive use in India: apprehension versus future intention among non-users and traditional method users. *Sex Reprod. Hlthcare*, **4**, 65–72.
Ram, U., Shekhar, C. and Chowdhury, B. (2014) Use of traditional contraceptive methods in India and its socio-demographic determinants. *Ind. J. Med. Res.*, **140**, 17–28.
Rizwan, S. A., Kankari, A., Roy, R. K., Upadhyay, R. P., Palanivel, C., Chellaiyan, V. G. and Babu, D. S. (2012) Effect of literacy on family planning practices among married women in rural south India. *Int. J. Med. Publ. Hlth*, **2**, 24–27.
Rousseau, J. and Mengersen, K. (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Statist. Soc.* B, **75**, 689–710.
Sääväla, M. (1999) Understanding the prevalence of female sterilization in rural South India. *Stud. Famly Planng*, **4**, 288–301.
Srikanthan, A. and Reid, R. L. (2008) Religious and cultural influences on contraception. *J. Obstetr. Gyn. Can.*, **30**, 127–137.
Tutz, G. (1991) Sequential models in categorical regression. *Computnl Statist. Data Anal.*, **11**, 275–295.