Childhood Obesity in Singapore: a Bayesian Nonparametric Approach

Mario Beraha [1], Alessandra Guglielmi [2], Fernando Andrés Quintana [3], Maria De Iorio [4], Johan Gunnar Eriksson [4] and Fabian Yap [5]

[1] ESOMAS, University of Torino, Torino, Italy

[2] Department of Mathematics, Politecnico di Milano, Milano, Italy

[3] Department of Statistics, Pontificia Universidad Católica de Chile, Santiago, Chile

[4] Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[5] Department of Paediatrics, KK Women's and Children's Hospital, Singapore

---

Address for correspondence: Alessandra Guglielmi, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci, 32 - 20133 Milano, Italy.

E-mail: alessandra.guglielmi@polimi.it.

Phone: .

Fax: .

Abstract: Overweight and obesity in adults are known to be associated with increased risk of metabolic and cardiovascular diseases. Obesity has now reached epidemic proportions, increasingly affecting children. Therefore, it is important to understand if this condition persists from early life to childhood and if different patterns

can be detected to inform intervention policies. Our motivating application is a study of temporal patterns of obesity in children from South Eastern Asia. Our main focus is on clustering obesity patterns after adjusting for the effect of baseline information. Specifically, we consider a joint model for height and weight over time. Measurements are taken every six months from birth. To allow for data-driven clustering of trajectories, we assume a vector autoregressive sampling model with a dependent logit stick-breaking prior. Simulation studies show good performance of the proposed model to capture overall growth patterns, as compared to other alternatives. We also fit the model to the motivating dataset, and discuss the results, in particular highlighting cluster differences. We have found four large clusters, corresponding to children sub-groups, though two of them are similar in terms of both height and weight at each time point. We provide interpretation of these clusters in terms of combinations of predictors.

Key words: clustering; longitudinal profiles; obesity development; covariate dependent priors

1 Introduction

Overweight and obesity are defined as abnormal or excessive fat accumulation that may impair health (WHO, 2022). It is well-known that overweight and obesity in adults are associated with higher risk of metabolic and cardiovascular diseases; see, for instance, Després et al. (2008), Fox et al. (2007) and Pi-Sunyer (2009). Furthermore, individuals who are obese and contracted COVID-19 are more likely to experience a

more severe course of illness (Gao et al., 2020).

Obesity is an epidemic, increasingly affecting children. In 2018, 18% of children in the United States were obese and approximately 6% were severely obese (Hales et al., 2018). Prevalence of obesity in children has increased from 4% in 1975 to over 18% in 2016 among children and adolescents aged 5-19 years [WHO, Accessed: 01-06-2021]; see also Cremaschi et al. (2021). Overweight or obesity in childhood is critical as it often persists into adulthood due to both physiological and behavioral factors, e.g. (i) adults diet based on energy-dense foods that are high in fat and sugars and (ii) adult physical inactivity due to the sedentary nature of many forms of work, changing modes of transportation, and increasing urbanization. Indeed, dietary composition and sedentary lifestyle have often been cited as main contributors to childhood obesity. Moreover, existing evidence suggest an important role of parents' socioeconomic status and maternal prenatal health indicators; see Cremaschi et al. (2021). Recent research suggests that susceptibility to metabolic disease may originate early in life. Different conditions in maternal uterus seem to influence metabolic health by altering glucose metabolism and body composition (Symonds et al., 2013; Godfrey et al., 2012). Moreover, increased adiposity has been observed in school-age children and infants (Nightingale et al., 2010; Whincup et al., 2005; Yajnik et al., 2002, 2003).

It is therefore important to understand whether obesity persists from early life to childhood and if different types of obesity development can be detected, as it will be later discussed in Section 4. For instance, Zhang et al. (2019) show that rates of change in Body Mass Index (BMI) at different childhood ages are differentially associated with adult obesity. Our motivating application is the study of obesity over time in a dataset of children from South East Asia (see Soh et al., 2014), with mea-

surements taken every six months from birth. In particular, we focus on height and weight. It is known that obesity might increase the risk of metabolic diseases, and that this risk is higher in Asian populations than in White Caucasian population (Misra and Khurana, 2011). In this work, we provide a flexible model for longitudinal vector responses such as children's height and weight to cluster children according to their growth patterns, i.e. the longitudinal trajectories, with the aim of uncovering different risk subgroups. This will inform the development of appropriate interventions. Our approach combines modelling growth curves with the flexibility of a covariate-dependent Bayesian nonparametric (BNP) mixture prior. The key idea is that we build clusters of individuals based not only on the shape of growth trajectories, but we also let the weights of our mixture prior depend on subject-specific covariates. As will be shown, this leads to clusters that are more homogeneous in terms of baseline features. Specifically, we assume a vector autoregressive (VAR) model to represent child growth, including subject-specific VAR parameters, after adjusting for covariates (both time-homogeneous and time-varying) available on children as well as mothers. Thus, clustering multivariate growth curves is equivalent to cluster the VAR parameters. We assume a covariate-dependent Bayesian nonparametric prior for the VAR parameters. A preliminary analysis shows that the lag 1 autoregression assumption is a reasonable approximation, with higher order lags implying no substantial gain in information. This is also a simpler and more parsimonious representation than alternatives such as a mixture of multivariate Gaussian distributions. The model also includes an overall time-dependent mean function.

In more detail, we assume the children-specific VAR coefficients to be independently distributed according to a truncated stick-breaking prior with weights that depend on baseline covariates. This construction induces a covariate-dependent prior on

the partition of the children in the sample. Moreover, it allows for potentially empty components, in which case the *number of clusters* is interpreted as the number of non-empty components in the stick-breaking representation, i.e. components to which at least one observation is assigned. The dependent stick-breaking prior adopted here can be seen as a finite-dimensional version of the logistic stick-breaking process described in Ren et al. (2011), which belongs to the family of covariate dependent random probability processes. See, for example, MacEachern (2000); Chung and Dunson (2009); Rodríguez and Dunson (2011); Müller et al. (2011); Park and Dunson (2010) and Quintana et al. (2022) for a review. A BNP approach is particularly appealing for the application under study, since comparison with alternative models shows that a parametric dependence structure is unable to fully capture the data complexity. We compare our approach also with a popular covariate-dependent prior, the linear dependent Dirichlet process (Linear-DDP); in this case, our prior shows better performance in terms of standard model metrics.

On the other hand, VAR models provide a flexible and powerful representation of longitudinal data, since they allow a straightforward representation of the covariance matrix of the data; see, for instance, Canova and Ciccarelli (2004) and Daniels and Pourahmadi (2002). Bayesian nonparametric methods have been successfully applied to VAR models in recent years. See Kalli and Griffin (2018) for such a model applied to single subject data, and Billio et al. (2019) and Kundu and Lukemire (2021) for multiple subject data. In Billio et al. (2019) the authors propose a Dirichlet process mixture of Normal-Gamma priors on the VAR autocovariance elements, as a Bayesian-Lasso prior. In Kundu and Lukemire (2021) the focus is on matrix-variate data, providing a class of nonparametric Bayesian VAR models, based on heterogeneous multi-subject data, that enable separate clustering at multiple scales, and

result in partially overlapping clusters. An alternative modeling approach is offered by longitudinal data techniques with fixed and/or random effects functions in time and latent stochastic processes (see, for instance, Li et al., 2010; Quintana et al., 2016, and references therein). The general context of dynamic models representation of longitudinal data with priors for the associated covariance matrices is illustrated in Daniels and Pourahmadi (2002), with the class of VAR models being a particular case. Instead Quintana et al. (2016) present a BNP model for longitudinal data that includes flexible mean functions and autoregressive covariance structures. Similarly to our proposal, their clustering is imposed on the autocorrelation structure across subjects, though cluster estimates are not their main inferential focus. Finally, Cremaschi et al. (2021) consider a more complex model in a similar framework, i.e. they propose a joint model for multiple growth markers and metabolic associations, which allows for data-driven clustering of the children and highlights metabolic pathways involved in child obesity. Unlike our approach, they model the longitudinal trajectory with a Gaussian Process and the metabolic associations with a Gaussian Graphical model, assuming a joint Bayesian nonparametric random effect distribution on the parameters characterizing the growth curves and the graph.

We introduce a Bayesian model to cluster obesity growth patterns whose key components are given by: (i) a VAR model, (ii) a covariate-dependent BNP prior for the VAR parameters driving the clustering, and (iii) the inclusion of fixed-time and time-varying covariates in the likelihood. An alternative to jointly model height and weight is the adoption of the unidimensional BMI curve as a response. However, when checking for children health growth, pediatricians usually focus on growth charts of both height and weight. This is how we proceed in this manuscript. We design a tailored efficient Gibbs sampling algorithm to perform posterior inference, that exploits

the recent results on logit stick-breaking priors by Rigon and Durante (2021).

The paper is structured as follows. We first describe the motivating application (Section 2), and then we introduce the finite mixture of VAR models and discuss its main features (Section 3). Next (Section 4) we present the results from the main application; we also include predictive goodness of fit measures to compare with alternative models. The paper concludes with a discussion in Section 5. Supplementary material file provides further plots, details on the Gibbs sampler algorithm for posterior simulation and on extensive simulation studies carried out to explore model performance and compare versus competitor models.

2 Motivating application

We focus on the analysis of obesity in children from Singapore, particularly on its evolution over time. As mentioned in the Introduction, it is relevant to understand whether obesity persists from early life to childhood. Such information is of particular relevance when designing intervention policies. We describe the main features of the data and introduce the research questions. Then, we present the results of an exploratory analysis carried out to highlight the main modelling challenges and further motivate our approach.

2.1 GUSTO Child growth dataset

We consider data from *the Growing Up in Singapore Towards healthy Outcomes* (GUSTO) study, which is one of the most carefully phenotyped parent-offspring cohorts with a particular focus on epigenetic observations; see Soh et al. (2014) for

description of subject participants and objectives of the cohort study. The data consist of measurements of child height (or length, depending on the child's age) in centimeters and weight in kilograms from periodic visits of 1139 children from birth to the age of seven. We consider only visits that occurred every six months, though during the first year of life, infants were examined every three months. This has been done since, after preliminary analysis, we have focused on autoregressive models for the bidimensional response. Such a class requires time units to be homogeneous. More specifically, the response vector $\boldsymbol{y}_{it} \in \mathbb{R}^2$ is given by the measurements of (*length*, *weight*) up to the 12th month of age ($t = 3$) and (*height*, *weight*) from the 18th month onwards ($t = 4, \ldots, 14$). Besides gender of the child, information is also available on the mother. However, the original sample includes missing observations: 77 subjects are discarded from the analysis, because only information on the first visit (i.e. right after birth) is available. Moreover, we discard children with less than two consecutive visits, and with missing baseline covariates. The number of children with missing baseline covariates is 217, while the number of children with less than two consecutive visits is 79. This leads to a final sample size of $N = 766$. Note that we keep children with missing responses, since in a Bayesian framework it is straightforward to impute these as part of the MCMC. To this end, we simulate the missing responses (percentages of missing height and weight are 3.1% and 0.14%, respectively) from their full conditional distribution at every iteration of the algorithm. See the MCMC algorithm in the Supplementary Material, Section 2.

The available baseline covariates are:

- *age*: mother's age. it ranges from 18 to 46 years.
- *parity*: number of previous pregnancies carried to a viable gestation by the

mother, ranging from 0 to 5. If parity equals to 0, the child is the first born.

- *OGTT fasting Pw26* (in what follows referred to as *OGTT fasting*): oral glucose tolerance test (OGTT) at 24th-26th week of pregnancy. It assumes values between 2.9 and 8.7 mg/dL. Mothers are tested after fasting for at least eight hours.

- *OGTT 2hour Pw26* (in what follows referred to as *OGTT 2h*): oral glucose tolerance test at 24th-26th week of pregnancy. It ranges from 2.9 to 15.1 mg/dL. Mothers are tested two hours after having assumed a glucose solution containing a high dose of sugar.

- *ppBMI*: pre-pregnancy body mass index of the mother. Values in the sample range from 14.6 to 41.3 Kg/m^2.

- *GA*: gestational age in weeks, i.e. the length of the pregnancy. It assumes values between 28 and 41.4.

- *sex*: gender of the child.

- Mother's *ethnicity*: Chinese, Malay or Indian with proportions consistent with the Singaporean population.

- Mother's highest *education*: categorical variable with three ordered levels. Level 1 corresponds to no education or primary school, level 2 corresponds either to primary school, GCE (Singapore-Cambridge general certificate of education (O-level)) or ITE NTC (institute of technical education, national technical certificate) and level 3 corresponds to university degree.

The main goal of the analysis is to understand differences in obesity growth patterns

among ethnic groups via the construction of clusters of individuals exhibiting different profiles. We are also interested in assessing the effect of gender, parity and gestational age of the children on the development of obesity (Tint et al., 2016). Gender, age and parity have been reported in the medical literature as associated to neonatal adiposity. Girls are known to have greater adiposity than boys even at birth (Simon et al., 2013; Fields et al., 2009; Rodríguez et al., 2004). Increasing parity is associated with increasing neonatal adiposity in Asians as well as in Western populations (Joshi et al., 2005; Catalano et al., 1995). Gestational age and postnatal age have also been shown to be associated with increasing weight and adiposity (Simon et al., 2013; Catalano et al., 1995). Other important factors relating to the mother are the results of the glucose tolerance test and pre-pregnancy body mass index, since metabolic diseases are heritable, though they do not necessarily lead to obesity (CDS, 2018); see also, for instance, Qasim et al. (2018). Since obesity might also be related to family nutritional habits, we include in the model *education* as a proxy for the family socioeconomic status.

In the next subsection, we present an exploratory data analysis, which will drive the choice of interactions between covariates to include in the model.

2.2 Exploratory data analysis

The three main ethnic groups in Singapore are Chinese, Malay and Indian. Their sample frequencies in the dataset, 56%, 26% and 18%, respectively, are consistent with the overall population distribution. In Figure 1 we plot the sample correlation of the numerical covariates. We find that the largest correlation (equal to 0.42) is between *OGTT fasting* and *OGTT 2h*, as expected. To better investigate the

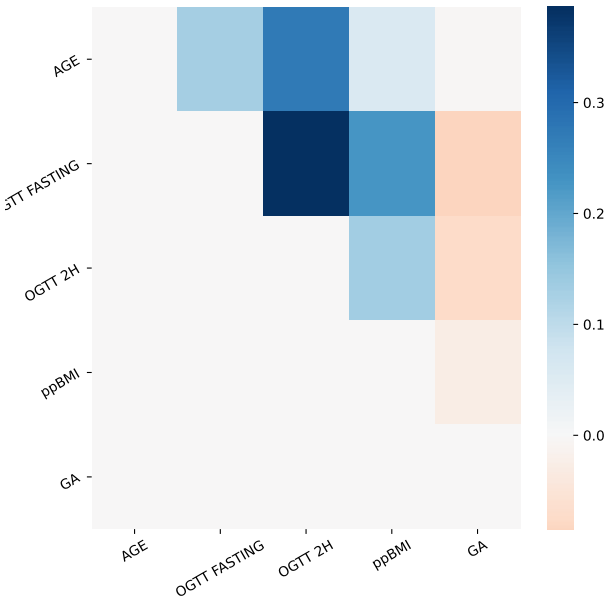


Figure 1: Sample correlation between numerical covariates in the Child Growth dataset.

relationship between categorical and continuous covariates, Figure 2 shows boxplots of the continuous covariates, stratified by each categorical covariate level. There appears to be a linear trend between *parity* and *age*, which is to be expected, and also between *parity* and *ppBMI*. Additionally, the distribution of mother's age is concentrated on smaller values for Malay and Indian ethnicity, compared to Chinese women. No other association is detectable between categorical and continuous covariates by visual inspection.

Figure 3 shows the scatterplots of the children's height (left) and weight (right) at lag 1, i.e. sample points (y_{it}, y_{it+1}) for all t and all subject i for both responses y. We identify two sub-groups in both plots, corresponding to newborns and infants (the group of datapoints on the left bottom corner of each panel) and older children. For

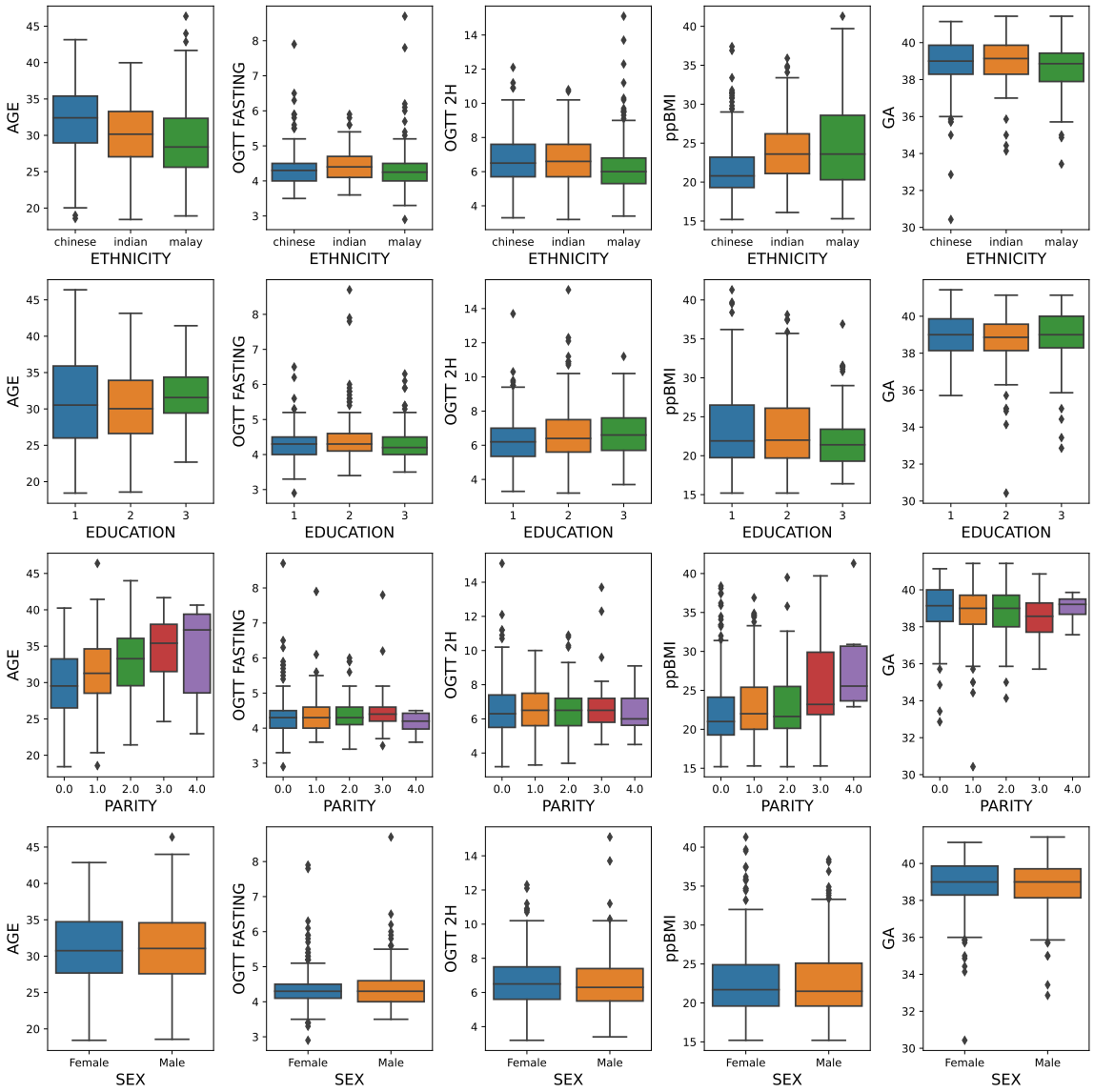


Figure 2: Boxplots of numerical variables (by column) for each level of the categorical variables (by row).

the latter the autoregressive assumption is very clear, while for the infant group, as expected, the linearity relationship is not strong, though it could be assumed as a first approximation.

In Section 1 of the Supplementary Material we show the unidimensional scatterplots of the responses (height and weight) at $t = 0, 1, 2$ versus the continuous covariates,

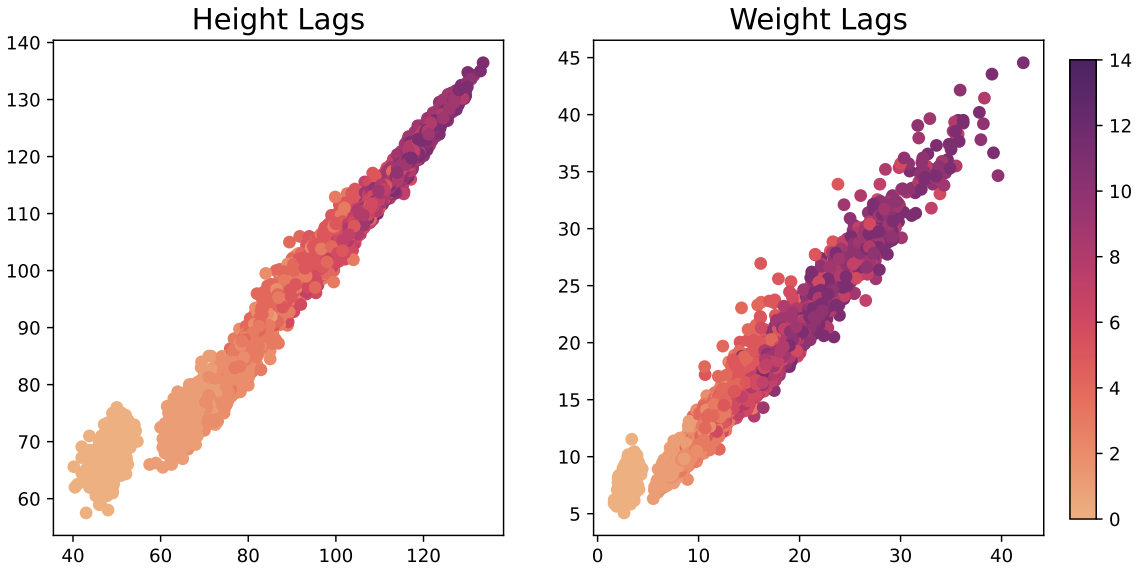


Figure 3: Scatterplots of Singapore children's height (left) and weight (right) at lag 1, i.e. of the sample points (y_{it}, y_{it+1}), for $t = 1, \ldots, T_i - 1$ and $i = 1, \ldots, N$ for each response y; the colors correspond to the age of the children as defined in the colorbar

which are useful to identify overall effects of these time-homogeneous (i.e. recorded at baseline) covariates, on the time-varying responses. For categorical covariates, we show the boxplots of responses stratified by level. Supplementary Material Figures 1 and 2 display a response pattern increasing with time, though there does not seem to be a clear dependence of weight and height on such covariates.

The lagged scatterplots in Figure 3 justify adopting a VAR model with lag 1 for the responses. Moreover, we include in the analysis the time-homogeneous covariates $\boldsymbol{z}_i$ and a function of time, $x_{it} = \sqrt{t}$, to account for a global growth trend over time. We have fitted different trends in a preliminary analysis, but the best fit was for $\sqrt{t}$, which is what we use here. No other time-varying covariate is available in the dataset. We also consider interaction terms between (i) the mother's highest education and age, and (ii) ethnicity and gender of the child. Finally, denoting by $X : Y$ the interaction

term between X and Y, Table 1 lists all the covariates included in the model:

	Covariate	Type		Covariate	Type
(1)	intercept		(9)	$education_2$*:age*	discrete
(2)	*age*	discrete	(10)	$education_3$*:age*	discrete
(3)	*parity*	discrete	(11)	*parity:age*	discrete
(4)	*OGTT fasting*	cont	(12)	*Indian*	binary (=1 if the mother is Indian and zero otherwise)
(5)	*OGTT 2h*	cont	(13)	*Malay*	binary (=1 if the mother is Malay and zero otherwise)
(6)	*ppBMI*	cont	(14)	*Male:Chinese*	binary (=1 for a male child born to a Chinese mother)
(7)	*GA*	cont	(15)	*Male:Indian*	binary (=1 for a male child born to an Indian mother)
(8)	$education_1$*:age*	discrete	(16)	*Male:Malay*	binary (=1 for a male child born to a Malay mother)

Table 1: Final model covariates and their data types: *cont* and *discrete* denote continuous and discrete numeric variables, respectively.

The baseline category for the categorical covariates corresponds to a female child born

to a Chinese mother. As a final pre-processing step, we standardize each numerical covariate at baseline by subtracting their sample mean and dividing by the sample standard deviation.

In summary, the Child Growth dataset used in the anaysis contains information on $N = 766$ children, $k = 2$ responses, $p = 1$ time-dependent covariate (i.e. $\sqrt{t}$) and a $q = 16$-dimensional design matrix for time-homogeneous covariates (including intercepts, interactions and dummy variables to represent categorical covariates).

3 The VAR model and the logit stick-breaking prior for the VAR parameters

Our motivating application requires the development of statistical methodology able to describe the evolution of a k-dimensional response vector $\boldsymbol{Y}_{it}$ for individuals i, $i = 1, \ldots, N$ recorded at discrete time points t, $t = 1, \ldots, T_i$, accounting for time-varying covariates $\boldsymbol{x}_{it}$ and time-homogeneous covariates $\boldsymbol{z}_i$, measured at the baseline. Motivated by the exploratory analysis above, for any $i = 1, \ldots, N$, we assume:

$$\boldsymbol{y}_{it} = \Phi_i \boldsymbol{y}_{it-1} + B\boldsymbol{x}_{it} + \Gamma \boldsymbol{z}_i + \boldsymbol{\varepsilon}_{it}, \quad \boldsymbol{\varepsilon}_{it} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \Sigma),\ t = 1, \ldots, T_i \tag{3.1}$$

where $\Phi_i = [\Phi_{ijl}]$ is a $k \times k$ matrix of autoregression coefficients, $\boldsymbol{x}_{it}$ is a $p-$dimensional vector of time-varying covariates, $\boldsymbol{z}_i$ is a $q-$dimensional vector of time-homogeneous covariates, $B = [b_{jl}]$ and $\Gamma = [\gamma_{jl}]$ are $k \times p$ and $k \times q$ matrices of regression coefficients, respectively. For ease of explanation, we vectorize matrices Φ_i, B and Γ. Specifically, denoting with $(\cdot)^T$ the transpose of a column vector, we introduce the

following notation

$$\begin{aligned}\varphi_i &= (\Phi_{i11}, \dots, \Phi_{i1k}, \Phi_{i21}, \dots, \Phi_{i2k}, \dots, \Phi_{ik1}, \dots, \Phi_{ikk})^T \\ \boldsymbol{b} &= (b_{11}, \dots, b_{1p}, b_{21}, \dots, b_{2p}, \dots, b_{k1}, \dots, b_{kp})^T \\ \boldsymbol{\gamma} &= (\gamma_{11}, \dots, \gamma_{1q}, \gamma_{21}, \dots, \gamma_{2q}, \dots, \gamma_{k1}, \dots, \gamma_{kq})^T\end{aligned}$$

so that φ_i, $\boldsymbol{b}$ and $\boldsymbol{\gamma}$ are vectors with k^2, $k \times p$ and $k \times q$ elements (vectorization of the matrices Φ_i, B, Γ, respectively). We assume $\boldsymbol{y}_{i0} = \mathbf{0}$, that is, conditionally to the remaining parameters, $\boldsymbol{y}_{i1}$ has a Gaussian distribution with mean $B\boldsymbol{x}_{i1} + \Gamma\boldsymbol{z}_i$. Alternatively, we could consider the responses at baseline as exogenous, or a different initial distribution could be specified. We assume that a priori $(\Phi_1, \dots, \Phi_N)$, $\boldsymbol{b}$, $\boldsymbol{\gamma}$ and Σ are independent. As random effect distribution, we assume a Bayesian nonparametric prior which depends on the baseline covariates. Specifically, we assume that

$$\Phi_i \mid z_i \overset{\text{ind}}{\sim} \sum_{h=1}^{H} w_h(\boldsymbol{z}_i)\delta_{\Phi_{0h}} \quad i = 1, \dots, N \tag{3.2}$$

and we impose a stick-breaking construction on the weights w_h. This implies that equation (3.2) defines a truncated stick-breaking prior with H support points $\{\Phi_{0h}\}$ and covariate-dependent weights summing up to 1. Similarly to Rigon and Durante (2021), we assume that the weights are generated via a logit stick-breaking construction, that is, $w_1(\boldsymbol{z}_i) = \nu_1(\boldsymbol{z}_i)$, and $w_h(\boldsymbol{z}_i) = \nu_h(\boldsymbol{z}_i)\prod_{l=1}^{h-1}(1 - \nu_l(\boldsymbol{z}_i))$ for $h = 1, \dots, H-1$, and $\nu_H(\boldsymbol{z}_i) = 1$. The dependence on the covariates $\boldsymbol{z}_i$ is introduced by assuming a logistic model for $\nu_h(\boldsymbol{z}_i)$:

$$\begin{aligned}\text{logit}(\nu_h(\boldsymbol{z}_i)) &= \boldsymbol{z}_i^T\boldsymbol{\alpha}_h, \quad h = 1, \dots, H-1 \\ \boldsymbol{\alpha}_h &\overset{\text{iid}}{\sim} \mathcal{N}_q(\mu_\alpha, \Sigma_\alpha), \quad h = 1, \dots H-1\end{aligned} \tag{3.3}$$

An equivalent formulation of (3.2) can be obtained by introducing auxiliary variables c_i's (usually referred to as cluster allocation indicators) such that

$$c_i \mid z_i, \boldsymbol{\alpha} \sim \text{Categorical}\left(\{1, \ldots, H\}; \boldsymbol{w}(\boldsymbol{z}_i)\right)$$

and letting $\Phi_i = \Phi_{0c_i}$. Availability of the c_i's allows us to make a fundamental distinction between mixture components and clusters. In the following, we refer to any of the Φ_{0h}'s as a *component*, while a *cluster* of observations is a component to which some observations are assigned to; see, for instance, Argiento and De Iorio (2022). The marginal prior (3.2) - (3.3) is represented by a finite, though large number of parameters, and can be regarded as the truncation of a dependent Bayesian nonparametric prior. We complete the prior specification with the marginal parametric prior distributions for $\boldsymbol{b}$, $\boldsymbol{\gamma}$ and Σ:

$$\boldsymbol{b} \sim \mathcal{N}_{kp}(\mathbf{0}, \Sigma_B), \qquad \boldsymbol{\gamma} \sim \mathcal{N}_{kq}(\mathbf{0}, \Sigma_\Gamma) \qquad \Sigma^{-1} \sim \mathcal{W}(\Sigma_0, \nu) \tag{3.4}$$

where $\mathcal{W}(\Sigma_0, \nu)$ denotes the Wishart distribution with expectation equal to $\nu\Sigma_0$ for $\nu > p - 1$. To obtain more robust inference, we assume a hierarchical prior for the φ_{0h}'s:

$$\varphi_{0h} \mid \varphi_{00}, V_0 \overset{\text{iid}}{\sim} \mathcal{N}_{k^2}(\varphi_{00}, V_0), \quad h = 1, \ldots, H \tag{3.5}$$

$$\varphi_{00}, V_0 \mid \varphi_{000}, \lambda, V_{00}, \tau_0 \sim \mathcal{NIW}(\varphi_{000}, \lambda, V_{00}, \tau_0) \tag{3.6}$$

In (3.6), $\mathcal{NIW}(\varphi_{000}, \lambda, V_{00}, \tau_0)$ denotes the normal-Inverse Wishart distribution, i.e. $V_0 \sim \mathcal{IW}(\tau_0, V_{00})$ and $\varphi_{00} \mid V_0 \sim \mathcal{N}(\varphi_{000}, \lambda^{-1}V_0)$, where $\mathcal{IW}(\tau_0, V_{00})$ denotes the inverse-Wishart distribution defined over the space of $k^2 \times k^2$ symmetric and positive definite matrices with mean $V_0/(\tau_0 - k^2 - 1)$.

Posterior inference is performed through a Gibbs sampler algorithm, as detailed in the Supplementary Material, Section 2. However, it is worth noting that the full-

conditional of the weights parameters $\{\boldsymbol{\alpha}_h\}$ in Equation (3.3) can be derived in closed-form with the introduction of auxiliary variables, using results in Polson et al. (2013) and Rigon and Durante (2021). The full conditional distributions of $\boldsymbol{b}$ and $\boldsymbol{\gamma}$ are derived as in a standard multivariate Bayesian linear regression models. The full conditionals of the atoms $\{\Phi_{0h}\}$ in the stick-breaking prior (3.2) are given in the blocked Gibbs sampling of Ishwaran and James (2001). The code is implemented in `C++` and linked to `Python` via `pybind11` (Jakob et al., 2017) and is publicly available at `https://github.com/mberaha/BNP-VAR.git`.

4 Child Growth data

We now present posterior results for the Child Growth dataset, detailing prior specification and inference. In the latter case we include a comparison between the proposed prior and the linear-DDP prior, as well as with a parametric counterpart of our model. Recall that the dataset contains information on $N = 766$ children with $k = 2$ responses, height and weight of the children over time.

4.1 Prior elicitation

Given the complexity of the model and the high-dimensionality of the dataset, prior elicitation needs to be carefully considered. Preliminary analysis shows that when the variances of the α_h's (see (3.3)) or of the atoms Φ_{0h}'s (see (3.5)) in the logit stick-breaking are large, then all the observations tend to be assigned to the same component. Moreover, the missing data simulation step has a strong impact on posterior inference. In particular, when using the vague prior described above, in the

initial iterations of the MCMC algorithm, typically large missing values are imputed (e.g. 10^5) since both Σ and $\{\Phi_{0h}\}$ would assume unusually large values. Sampled values for all the other parameters are affected, leading to a poor fit. Hence, the use of an uninformative prior is not advisable, as it causes poor mixing and slow convergence of the chain. This is a common situation in complex hierarchical models when non-informative priors are adopted in lower levels.

As such, we opt for informative priors. To set the hyperparameters in the hierarchical marginal prior in (3.5)-(3.6), we adopt an empirical Bayes type of approach and obtain the maximum likelihood estimator of a vector autoregressive model:

$$\boldsymbol{y}_{it} \mid \boldsymbol{y}_{it-1} \sim \mathcal{N}(\Phi \boldsymbol{y}_{it-1}, \Sigma), \quad t = 1, \dots T-1,\ i = 1, \dots, N \tag{4.1}$$

which corresponds to (3.1) when B and Γ are set to zero (their prior expected value) and $H = 1$. We fit (4.1) using only subjects with no missing responses. Let $\widehat{\Phi}$, $\widehat{\Sigma}$ denote the maximum likelihood estimator for Φ and Σ respectively. We fix $\Phi_{000} = \widehat{\Phi}$, $\lambda = 1$, and select (V_{00}, τ) in (3.6) so that $\mathbb{E}[V_0] = I$ and $\text{Var}[\{V_0\}_{ii}] = 1.5$. Similarly, we fix Σ_0 and ν in (3.4) so that $\mathbb{E}[\Sigma] = \widehat{\Sigma}$ and $\text{Var}[\{\Sigma_{ii}\}] = 10$.

The variance hyperparameter Σ_α in (3.3) also has an important effect on posterior inference. To set this quantity, we look at the prior distribution of the number of clusters (i.e. *occupied components*) and of the size of the largest cluster. To this end, we perform Monte Carlo simulations. Specifically, we fix the number of components H in the stick-breaking prior equal to 50, set $\Sigma_\alpha = \sigma_\alpha^2 I$, and simulate $\alpha_1, \dots, \alpha_{H-1}$ from (3.3) with $\mu_\alpha = \mathbf{0}$. Then, for each of the $N = 766$ subjects, we compute the associated weights $\boldsymbol{w}(\boldsymbol{z}_i)$ from the logit stick-breaking process, using observed covariates $\boldsymbol{z}_i$, and allocate each subject to one of the H components with probability given by the weights $\boldsymbol{w}(\boldsymbol{z}_i)$. The above procedure is repeated independently for $M = 10,000$ iterations and

we record the number of clusters and the size of the largest cluster. Figure 4 shows the distributions obtained from the Monte Carlo simulation. As σ_α^2 increases, the number of clusters shrinks to 1 and the size of the largest cluster increases accordingly. Hence, we fix $\sigma_\alpha^2 = 5$ so that a priori we should expect approximately $4 - 7$ clusters. Finally, we assume $\mu_\alpha = \mathbf{0}$, $\Sigma_B = I_2$ and $\Sigma_\Gamma = I_{18}$ (see (3.4)); recall that all continuous covariates are standardized.

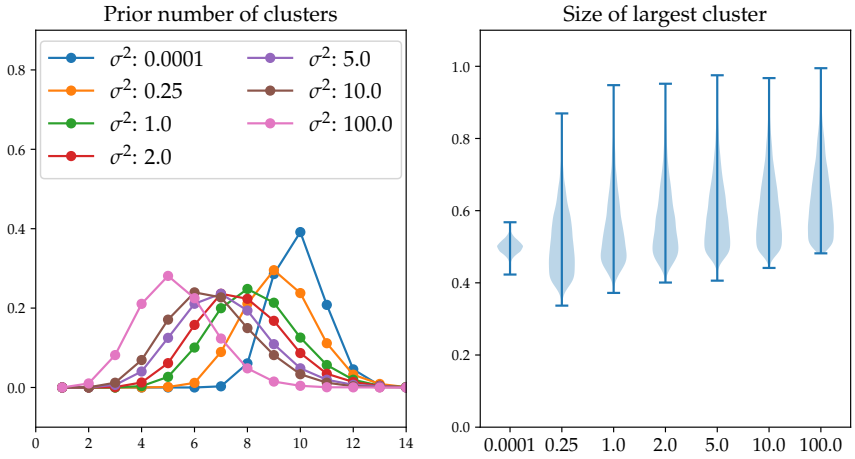


Figure 4: Prior distribution of the number of clusters (left panel) and of the size of the largest cluster as percentage of the whole dataset (right panel), for different values of σ_α.

4.2 Posterior inference results

We apply the model described in the previous section to the Child Growth dataset with hyperparameters set as above. Recall that the model includes $p = 1$ time-dependent covariate (that is, $\sqrt{t}$) and a $q = 16$-dimensional design matrix for time-homogeneous covariates (including intercepts, interactions and dummy variables to represent categorical covariates). We run the MCMC algorithm for $100,000$ iterations, discarding the first 50,000 as burn-in and thinning every 10 iterations, obtaining a

final sample size of 5,000 iterations. The chain has reached convergence as we can see from Figure 3 in the Supplementary Material file.

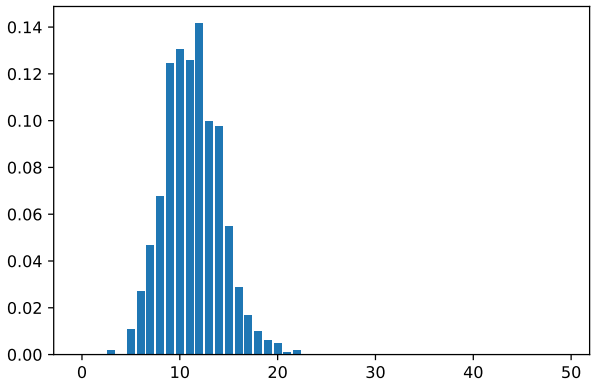


Figure 5: Child Growth dataset: posterior distribution of the number of clusters.

Figure 5 shows the posterior distribution of the number of clusters, i.e. of *occupied* parametric components, that is clearly centered around 10-12 clusters. However, interpreting these as the "number of distinct profiles" in the $\boldsymbol{y}$'s may be misleading. Recall that we have specified a covariate-dependent prior for the random partition of patients. Indeed, some clusters can be essentially identical when looking at the response trajectories, but different in terms of baseline features. As a point estimate of the latent partition, we choose the one that minimizes the posterior expectation of Binder's loss function (Binder, 1978). This loss function selects the clustering allocation that minimises the distance with the true probability of co-clustering between each pair of subjects, by assigning a cost b when two elements are wrongly clustered together and cost a when two elements are erroneously assigned to different clusters. We assume equal misclassification costs, i.e. $a = b$. The estimated partition consists of seven clusters, of which only four contain at least 15 observations. As it is usual in this type of literature (see, e.g. Page and Quintana, 2015), we focus only on these

four large clusters. In Figure 6 we display the response trajectories clustered according to the estimated partition. Note that the fourth cluster (bottom row) consists of subjects with at most three visits, except for one single subject with four visits. For this reason, we do not discuss this cluster. Figure 6 shows the time trajectories for patients' height (first column), weight (second column) and BMI. The third row in Figure 6 shows that this cluster contains children with lower weight, and consequently lower BMI than the other two clusters.

As already mentioned, the main three clusters could differ either in the responses or in the covariate patterns (or both). To better understand what discriminates the three main clusters, we perform homogeneity tests for the equality in distribution of both responses and covariates in the different clusters. The results should be considered as a descriptive tool. In particular, for the responses we consider the data on both height and weight at each visit separately and test equality of the distributions for each pair of clusters. For each of the covariates, we test the equality of their distributions in each possible pair of clusters. For the response variables and continuous covariates, we employ the Kolmogorov-Smirnov (KS) test for equality in distribution and the Pearson's chi- squared test of homogeneity for categorical covariates. Table 2 reports the p-values associated to the KS test for responses, while Figure 7 shows the cluster specific empirical distribution of covariates. From Table 2 and Figure 7, it is clear that clusters 2 and 3 (second and third rows in Figure 6, respectively) are similar in terms of both responses at each time point. However, Figure 7 (bottom row) suggests that the three main clusters cannot be *explained* only in terms of *ethnicity*, even though cluster 3 contains almost exclusively Chinese children.

Because the model-based clustering of children's multivariate trajectories is driven by

the clustering of the VAR parameters Φ_i's, as a further assessment of differences in cluster responses, we estimate those parameters. Since the Φ_i's are not identifiable due to label-switching of its mixture prior, we use an ad-hoc estimation procedure that has become standard. See Section 1 of the Supplementary Material. Those figures show that the estimates by cluster are different, including clusters 2 and 3.

We now discuss posterior inference on the two parameters contained in B, i.e. the regression parameters for the square root of time t for the two responses; see (3.1). The posterior means are $5.55 and 0.96$, respectively, with marginal standard deviations 0.02 and 0.01, thus indicating a non-negligible growth trend for both height and weight, as expected. Figure 8 displays posterior credible intervals for all the parameters in Γ defined in (3.1), that is, the regression coefficients corresponding to time-homogeneous covariates. The reference group for the categorical covariates is a Chinese female child. Covariates such as *OGTT 2h*, *ppBMI*, the interaction between education and age, ethnicity (Malay) and the interaction between gender and ethnicity have the strongest effects on height. On the other hand, *parity*, *OGTT 2h*, *ppBMI*, the interaction between education and age (but only the second level of education) and the interaction between gender and ethnicity have a strong association with weight. It is clear from Figure 8 that most of the posterior mass of the marginal distribution of regression coefficient *ethnicity* is concentrated on positive values. Correcting for the autoregressive effect, we see that *ethnicity* might impact obesity as Indian and Malay children are characterized by a larger posterior expected weight, combined in some cases with a lower posterior expected height. Moreover, also correcting for the autoregressive effect, our analysis shows that the posterior expected height of a Chinese male child is larger than the reference (Chinese female child). Similar comments can be made, for instance, regarding Indian male children being smaller than Indian

female children, and so on.

Mother's age and gestational age do not have a strong effect on the child's height and weight, though this might be due to the fact that these variables are associated with ethnicity (see Figure 2). It is known from the literature that increasing parity is associated with increasing neonatal adiposity in Asian and Western populations (see Tint et al., 2016); this is confirmed by the marginal posterior distribution of the parameter corresponding to the effect of *parity* on weight in Figure 8.

The time-homogeneous covariates z_i play also a key role in defining the stick-breaking prior as seen from (3.3). To assess if the proposed covariate-driven stick-breaking prior provides significant advantages over more standard models, we compare it with three possible competitors. The first one is the parametric version of our model obtained by setting $H = 1$. The second model assumes a truncated Dirichlet process as a prior for Φ_i's, with $H = 50$, similarly to what is done in the simulation study in the Supplementary Material file. The third competing prior assumes that the Φ_i's take into account information from the time-homogeneous covariates through the atoms Φ_{0h}'s. Specifically, the prior for Φ is specified as in (3.2), but for each $h = 1, \dots, H$ we define a matrix $\Omega_h \in \mathbb{R}^{k^2 \times q}$ and we let $vec(\Phi_{0h}(\boldsymbol{z}_i)) =: \varphi_{0h}(\boldsymbol{z}_i) = \Omega_h \boldsymbol{z}_i$. The weights $\boldsymbol{w}$ in (3.2) do not depend on the value of $\boldsymbol{z}_i$ (i.e., $w_h(\boldsymbol{z}_i) = w_h$) and follow a truncated Dirichlet process prior with $H = 50$. This model can be seen as a finite dimensional approximation of the Linear-DDP in De Iorio et al. (2004).

For all the models, we match the prior for B, Γ, Σ and, when possible also H and the marginal prior distribution of Φ_{0h}. For the Linear-DDP we assume that the vectorization of the Ω_h's are independent and identically distributed multivariate Gaussian random variables with mean zero and identity covariance matrix. Since the

full conditional distribution of the Ω_h's in the case of the Linear-DDP prior does not belong to a known parametric family, we update them via an adaptive Metropolis Hastings (Andrieu and Thoms, 2008) step.

The different models are compared using the widely applied information criterion (WAIC, Watanabe, 2013). Higher values of WAIC correspond to better predictive performance. We marginalise the missing values from the predictive distribution of the response trajectory and consider just the marginal predictive distribution for the non-missing values. We found that WAIC is equal to -3.4×10^6 for the Linear-DDP, -6.7×10^5 for the parametric model, -3.9×10^5 for the DP model and -3.4×10^5 for our model, confirming that our model performs better than the competitors. Moreover, we report that the MCMC algorithm for the Linear-DDP requires a much larger number of burn-in iterations (10^5 vs. 10^4) than the other models to reach satisfactory convergence, and that the posterior expected number of cluster in the Linear-DDP is around 42. It is then clear that (i) assuming linear dependence of the fixed-time covariates in the autoregressive parameters matrices Φ_i does not give good predictive fit (or at least not better than our model), and that (ii) adding covariate information in the stick-breaking prior improves the prediction performance. This is in line with the fact that models with covariate-dependent weights are more flexible than models that assume dependence only in the locations of a collection of random distributions.

5 Conclusions

Obesity is an epidemic, increasingly affecting children. Overweight or obesity in childhood may be critical as they often persist into adulthood due to both physiological and behavioral factors. The aim of this manuscript is to gain a better understanding of the factors affecting childhood obesity patterns.

We develop a Bayesian nonparametric VAR joint model for height and weight profiles and fit it to data from a Singaporean cohort study. One key aspect behind our modeling strategy is to cluster the joint time-evolving profiles using available covariate information. The model assumes a logit stick-breaking construction that can accommodate covariate dependent weights in the mixture model. This allows us to relate certain baseline features of children, such as gender or ethnicity, to obesity patterns. Ethnic differences in obesity are of interest as they could be due to genetic factors, dietary habits, cultural or socioeconomic factors. The analysis allows us to identify children sub-groups characterized by differences in time trajectories, covariates or both. Our discussion focused on three of the largest clusters detected. They differ most in terms of OGTT 2h. This is screening for maternal gestational diabetes, which is typically associated to their offspring's overweight and obesity risk. Cluster 3 is characterized by lower weight and slightly lower height than others, while cluster 1 differs in age from the others. Moreover, the estimates of Φ_i's in the Supplementary Material file in the main three clusters exhibit clear differences.

Posterior inference is carried out via an efficient sampling scheme that exploits recently developed results on logit stick-breaking priors. The results obtained are compared against competitor models, and we find that our approach provides superior

performance as measured by model choice criteria such as the WAIC.

An interesting characteristic of our model is that, though it clusters obesity patterns, when we characterize the estimated clusters we need to account for "number of distinct profiles" in the responses, as well as for the fact that the random partition of the subjects is covariate-dependent. In fact, some of the estimated clusters are similar when looking at the response trajectories, but different in terms of the covariate patterns. This is one of the most appealing advantages of our model (and all models with covariate-dependent prior for the random partition), as it allows for greater flexibility and interpretability.

Given the complexity of the data, the model is composed of four main components: an AR structure in the likelihood, the mean temporal trend, the interactions in the linear regression term, and the BNP prior for clustering. Moreover, the MCMC scheme exploits efficient computation for the logit stick-breaking prior developed by Rigon and Durante (2021), which ensures scalability of the proposed approach.

Acknowledgements

This work was partially funded by grants FONDECYT 1180034 and 1220017.

References

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and computing*, **18**(4), 343–373.

Argiento, R. and De Iorio, M. (2022). Is infinity that far? a Bayesian nonparametric perspective of finite mixture models. *The Annals of Statistics*, **50**(5), 2641–2663.

Billio, M., Casarin, R., and Rossini, L. (2019). Bayesian nonparametric sparse VAR models. *Journal of Econometrics*, **212**(1), 97–115.

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika*, **65**(1), 31–38.

Canova, F. and Ciccarelli, M. (2004). Forecasting and turning point predictions in a Bayesian panel VAR model. *Journal of Econometrics*, **120**(2), 327–359.

Catalano, P., Drago, N., and Amini, S. (1995). Factors affecting fetal growth and body composition. *Am J Obstet Gynecol.*, **172**(5), 1459–63.

CDS (2018). Centers for disease control and prevention - behavior, environment, and genetic factors all have a role in causing people to be overweight and obese. URL https://www.cdc.gov/genomics/resources/diseases/obesity/index.htm. Accessed: 19-01-2018.

Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, **104**(488), 1646–1660.

Cremaschi, A., De Iorio, M., Kothandaraman, N., Yap, F., Tint, M. T., and Eriksson, J. (2021). Integrating metabolic networks and growth biomarkers to unveil potential mechanisms of obesity. *arXiv preprint arXiv:2111.06212*.

Daniels, M. J. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, **89**(3), 553–566.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, **99**(465), 205–215.

Després, J.-P., Lemieux, I., Bergeron, J., Pibarot, P., Mathieu, P., Larose, E., Rodés-Cabau, J., Bertrand, O. F., and Poirier, P. (2008). Abdominal obesity and the metabolic syndrome: Contribution to global cardiometabolic risk. *Arteriosclerosis, Thrombosis, and Vascular Biology*, **28**(6), 1039–1049.

Fields, D., Krishnan, S., and Wisniewski, A. (2009). Sex differences in body composition early in life. *Gend Med.*, **6**(2), 369–75.

Fox, C. S., Massaro, J. M., Hoffmann, U., Pou, K. M., Maurovich-Horvat, P., Liu, C.-Y., Vasan, R. S., Murabito, J. M., Meigs, J. B., Cupples, L. A., D'Agostino, R. B., and O'Donnell, C. J. (2007). Abdominal visceral and subcutaneous adipose tissue compartments. *Circulation*, **116**(1), 39–48.

Gao, F., Zheng, K. I., Wang, X.-B., Sun, Q.-F., Pan, K.-H., Wang, T.-Y., Chen, Y.-P., Targher, G., Byrne, C. D., George, J., et al. (2020). Obesity is a risk factor for greater covid-19 severity. *Diabetes care*, **43**(7), e72–e74.

Godfrey, K. M., Haugen, G., Kiserud, T., Inskip, H. M., Cooper, C., Harvey, N. C. W., Crozier, S. R., Robinson, S. M., Davies, L., the Southampton Women's Survey Study Group, and Hanson, M. A. (2012). Fetal liver blood flow distribution: Role in human developmental strategy to prioritize fat deposition versus brain development. *PLOS ONE*, **7**(8), 1–7. doi: 10.1371/journal.pone.0041759.

Hales, C. M., Fryar, C. D., Carroll, M. D., Freedman, D. S., and Ogden, C. L. (2018). Trends in obesity and severe obesity prevalence in us youth and adults by sex and age, 2007-2008 to 2015-2016. *Jama*, **319**(16), 1723–1725.

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

Jakob, W., Rhinelander, J., and Moldovan, D. (2017). pybind11 – seamless operability between c++11 and python. https://github.com/pybind/pybind11.

Joshi, N., Kulkarni, S., Yajnik, C., Joglekar, C., Rao, S., Coyaji, K., H.G., L., Rege, S., and Fall, C. (2005). Increasing maternal parity predicts neonatal adiposity: Pune Maternal Nutrition Study. *Am J Obstet Gynecol*, **Sep;193**(3 Pt 1), 783–9.

Kalli, M. and Griffin, J. E. (2018). Bayesian nonparametric vector autoregressive models. *Journal of econometrics*, **203**(2), 267–282.

Kundu, S. and Lukemire, J. (2021). Non-parametric Bayesian vector autoregression using multi-subject data. *arXiv preprint arXiv:2111.08743*.

Li, Y., Lin, X., and Müller, P. (2010). Bayesian inference in semiparametric mixed models for longitudional data. *Biometrics*, **66**(1), 70–78. ISSN 0006-341X.

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

Misra, A. and Khurana, L. (2011). Obesity-related non-communicable diseases: South asians vs white caucasians. *International journal of obesity*, **35**(2), 167–187.

Müller, P., Quintana, F., and Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, **20**(1), 260–278.

Nightingale, C. M., Rudnicka, A. R., Owen, C. G., Cook, D. G., and Whincup, P. H. (2010). Patterns of body size and adiposity among UK children of South Asian, black African–Caribbean and white European origin: Child Heart And health Study in England (CHASE Study). *International Journal of Epidemiology*, **40**(1), 33–44.

Page, G. L. and Quintana, F. A. (2015). Predictions based on the clustering of heterogeneous functions via shape and subject-specific covariates. *Bayesian Analysis*, **10**(2), 379–410.

Park, J.-H. and Dunson, D. B. (2010). Bayesian generalized product partition model. *Statistica Sinica*, pages 1203–1226.

Pi-Sunyer, X. (2009). The medical risks of obesity. *Postgraduate medicine*, **121**(6), 21–33.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, **108**(504), 1339–1349. doi: 10.1080/01621459.2013.829001.

Qasim, A., Turcotte, M., De Souza, R., Samaan, M., Champredon, D., Dushoff, J., Speakman, J., and Meyre, D. (2018). On the origin of obesity: identifying the biological, environmental and cultural drivers of genetic risk among human populations. *Obesity reviews*, **19**(2), 121–149.

Quintana, F. A., Mueller, P., Jara, A., and MacEachern, S. N. (2022). The dependent Dirichlet process and related models. *Statistical Science*, **37**(1), 24–41. ISSN 0883-4237.

Quintana, F. A., Johnson, W. O., Waetjen, L. E., and B. Gold, E. (2016). Bayesian nonparametric longitudinal data analysis. *Journal of the American Statistical Association*, **111**(515), 1168–1181.

Ren, L., Du, L., Carin, L., and Dunson, D. B. (2011). Logistic stick-breaking process. *Journal of Machine Learning Research*, **12**(1).

Rigon, T. and Durante, D. (2021). Tractable Bayesian density regression via logit stick-breaking priors. *Journal of Statistical Planning and Inference*, **211**, 131–142.

Rodríguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, **6**, 145–177. doi: 10.1214/11-BA605.

Rodríguez, G., Samper, M. P., Ventura, P., Moreno, L. A., Olivares, J. L., and Pérez-González, J. M. (2004). Gender differences in newborn subcutaneous fat distribution. *European journal of pediatrics*, **163**(8), 457—461. ISSN 0340-6199. doi: 10.1007/s00431-004-1468-z.

Simon, L., Borrego, P., Darmaun, D., Legrand, A., Rozé, J., and Chauty-Frondas, A. (2013). Effect of sex and gestational age on neonatal body composition. *Br J Nutr.*, **109**(6), 1105–8.

Soh, S.-E., Tint, M. T., Gluckman, P. D., Godfrey, K. M., Rifkin-Graboi, A., Chan, Y. H., Stünkel, W., Holbrook, J. D., Kwek, K., Chong, Y.-S., et al. (2014). Cohort

profile: Growing up in singapore towards healthy outcomes (gusto) birth cohort study. *International journal of epidemiology*, **43**(5), 1401–1409.

Symonds, M., Mendez, M., Meltzer, H., Koletzko, B., Godfrey, K., Forsyth, S., and van der Beek, E. (2013). Early Life Nutritional Programming of Obesity: Mother-Child Cohort Studies. *Ann Nutr Metab*, **62**(2), 137–145.

Tint, M. T., Fortier, M. V., Godfrey, K. M., Shuter, B., Kapur, J., Rajadurai, V. S., Agarwal, P., Chinnadurai, A., Niduvaje, K., Chan, Y.-H., et al. (2016). Abdominal adipose tissue compartments vary with ethnicity in asian neonates: Growing up in singapore toward healthy outcomes birth cohort study. *The American journal of clinical nutrition*, **103**(5), 1311–1317.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, **14**(Mar), 867–897.

Whincup, P. H., Gilg, J. A., Owen, C. G., Odoki, K., Alberti, K. G. M. M., and Cook, D. G. (2005). British south asians aged 13–16 years have higher fasting glucose and insulin levels than europeans. *Diabetic Medicine*, **22**(9), 1275–1277.

WHO (2022). World Health Organization - Obesity and overweigth. URL https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight. Accessed: 11-01-2022.

Yajnik, C. S., Lubree, H. G., Rege, S. S., Naik, S. S., Deshpande, J. A., Deshpande, S. S., Joglekar, C. V., and Yudkin, J. S. (2002). Adiposity and Hyperinsulinemia in Indians Are Present at Birth. *The Journal of Clinical Endocrinology & Metabolism*, **87**(12), 5575–5580.

Yajnik, C. S., Fall, C. H. D., Coyaji, K. J., Hirve, S. S., Rao, S., Barker, D. J. P., Joglekar, C., and Kellingray, S. (2003). Neonatal anthropometry: the thin–fat Indian baby. The Pune Maternal Nutrition Study. *International Journal of Obesity*, **27**(2), 173–180.

Zhang, T., Whelton, P. K., Xi, B., Krousel-Wood, M., Bazzano, L., He, J., Chen, W., and Li, S. (2019). Rate of change in body mass index at different ages during childhood and adult obesity risk. *Pediatric obesity*, **14**(7), e12513.

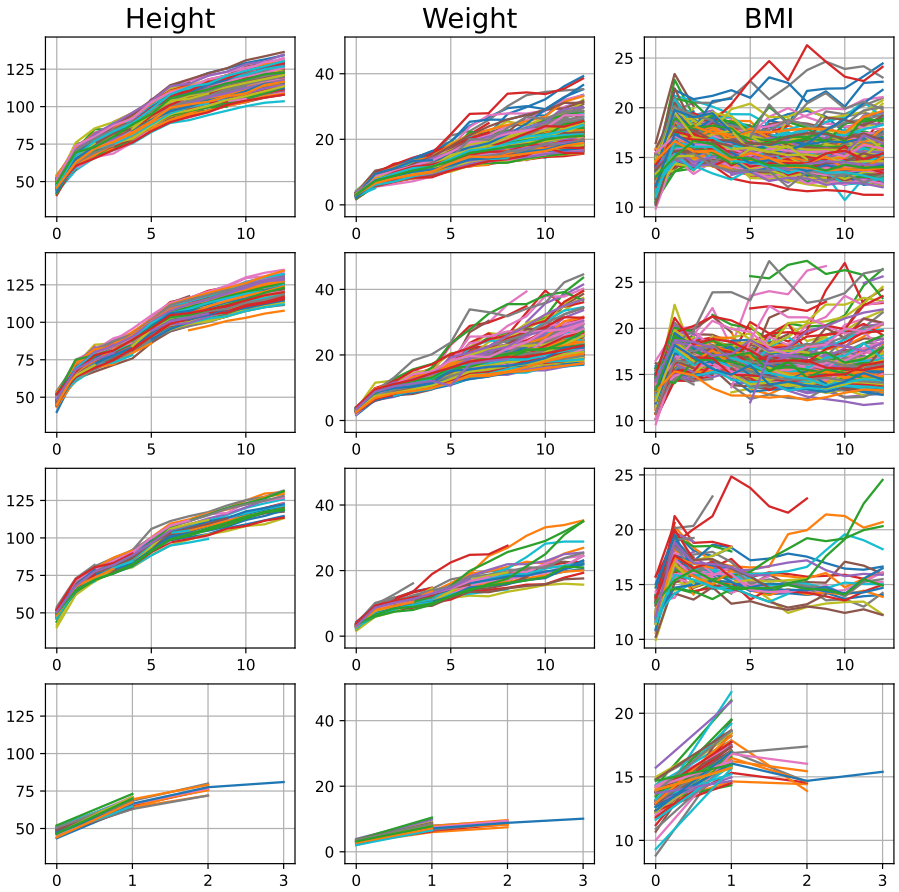
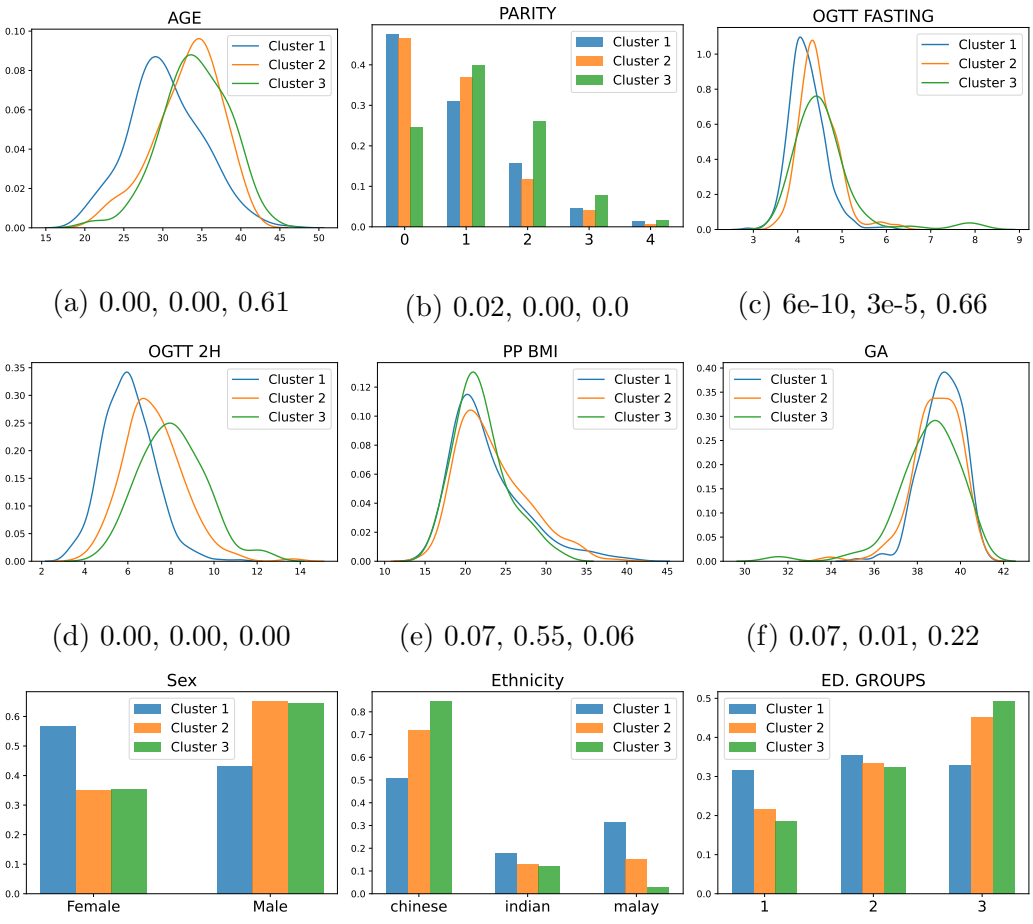


Figure 6: Subject-specific trajectories of height (first column), weight (second column) and BMI (third column) by estimated cluster (by row). The figure reports only the four largest clusters out of the seven estimated.

	Height			Weight		
Clusters	(1, 2)	(1, 3)	(2, 3)	(1, 2)	(1, 3)	(2, 3)
$t = 1$	**0.023**	**0.000**	**0.025**	**0.002**	**0.296**	0.606
$t = 2$	**0.000**	**0.023**	0.999	**0.000**	**0.000**	0.785
$t = 3$	**0.000**	**0.003**	0.797	**0.000**	**0.013**	0.815
$t = 4$	**0.000**	**0.000**	**0.000**	**0.000**	0.253	0.620
$t = 5$	**0.046**	**0.004**	**0.044**	**0.000**	0.197	0.386
$t = 6$	**0.000**	0.051	0.701	**0.000**	0.241	0.254
$t = 7$	**0.003**	0.113	0.878	**0.000**	0.431	0.375
$t = 8$	**0.000**	0.072	0.733	**0.000**	0.210	0.718
$t = 9$	**0.000**	0.106	0.984	**0.000**	0.196	0.715
$t = 10$	**0.000**	0.112	0.869	**0.000**	0.341	0.717
$t = 11$	**0.000**	0.213	0.726	**0.000**	0.244	0.854
$t = 12$	**0.000**	0.165	0.877	**0.000**	0.125	0.932
$t = 13$	**0.000**	0.179	0.993	**0.000**	**0.042**	0.811

Table 2: P-values of the homogeneity tests for the equality in distribution at every visit for each pair of clusters, considering height and weight. Bold numbers correspond to p-values lower than 5%



(a) 0.00, 0.00, 0.61 (b) 0.02, 0.00, 0.0 (c) 6e-10, 3e-5, 0.66

(d) 0.00, 0.00, 0.00 (e) 0.07, 0.55, 0.06 (f) 0.07, 0.01, 0.22

(g) 0.00, 0.00, 0.98 (h) 0.00, 0.00, 0.00 (i) 0.00, 0.00, 0.40

Figure 7: Empirical distribution of the covariates in each cluster. The three numbers below each plot represent the p-values for the homogeneity tests for covariates in clusters (1, 2), (1, 3) and (2, 3), respectively.

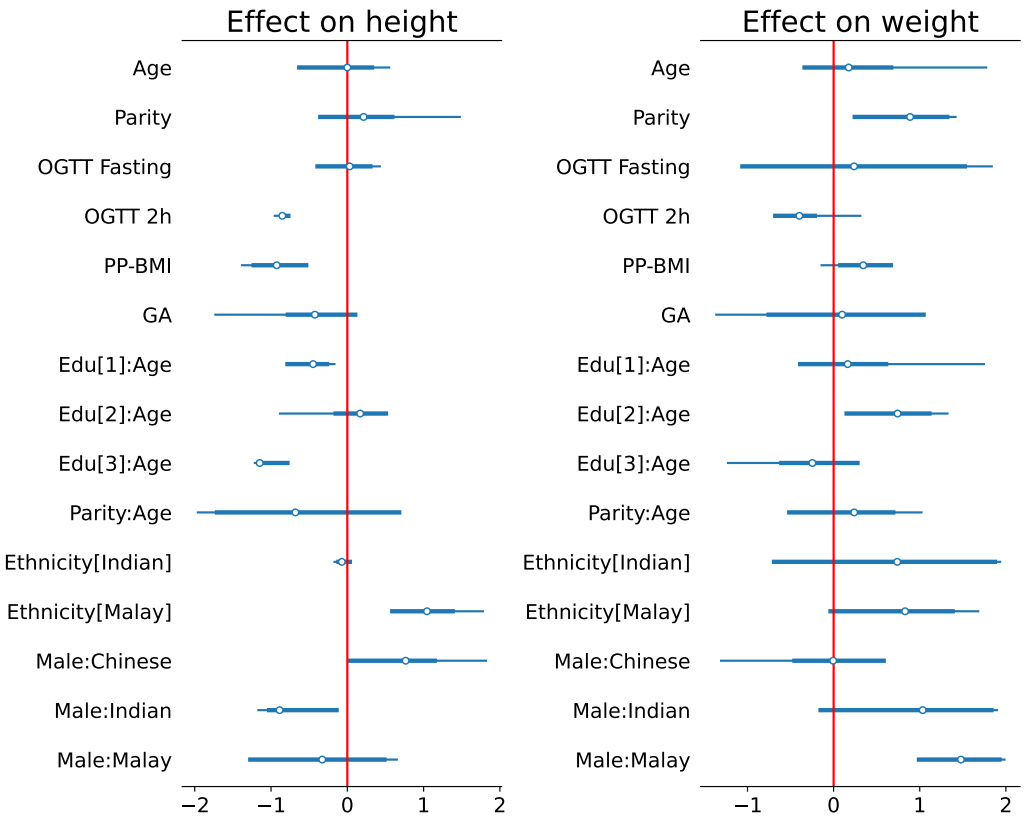


Figure 8: Posterior credible intervals of the regression coefficients in Γ for the height (left plot) and weight of the children (right plot). Thin lines correspond to 95% credible intervals, while thick lines to 80% credible intervals.