Marco Molinari*, Maria de Iorio, Nishi Chaturvedi, Alun Hughes and Therese Tillin

Modelling ethnic differences in the distribution of insulin resistance via Bayesian nonparametric processes: an application to the SABRE cohort study

https://doi.org/10.1515/ijb-2019-0108
Received September 27, 2019; accepted July 27, 2020; published online August 31, 2020

Abstract: We analyse data from the Southall And Brent REvisited (SABRE) tri-ethnic study, where measurements of metabolic and anthropometric variables have been recorded. In particular, we focus on modelling the distribution of insulin resistance which is strongly associated with the development of type 2 diabetes. We propose the use of a Bayesian nonparametric prior to model the distribution of Homeostasis Model Assessment insulin resistance, as it allows for data-driven clustering of the observations. Anthropometric variables and metabolites concentrations are included as covariates in a regression framework. This strategy highlights the presence of sub-populations in the data, characterised by different levels of risk of developing type 2 diabetes across ethnicities. Posterior inference is performed through Markov Chains Monte Carlo (MCMC) methods.

Keywords: cluster analysis; dirichlet process; gibbs sampling; metabolomics; SABRE study.

1 Introduction

The global epidemic of type 2 diabetes disproportionately affects non-European ethnic groups. South-Asians (from the Indian subcontinent) form the largest ethnic minority group in the UK with prevalence of diabetes in South-Asians estimated to be 2–4 times higher than that of the general population [1]. Africans-Caribbean in the UK, although fewer in number, are also at greater risk of developing type 2 diabetes, with prevalence also estimated at 2–4 times that of the general UK population [1].

The causal mechanisms underlying development of type 2 diabetes remain poorly understood, and no study has yet conclusively explained the reasons for the excess risk of diabetes experienced by South-Asian and African-Caribbean populations, suggesting that complex metabolic disturbances may underlie the ethnic differences [2]. Insulin resistance is a frequent precursors of type 2 diabetes in all populations and can be measured non-invasively using indices such as the Homeostasis Model Assessment (HOMA IR), which can be calculated from fasting blood glucose and insulin levels [3].

The main purpose of this work is to explore potential mechanisms underlying the marked ethnic differences in insulin resistance (Figure 1). A wide range of metabolic and phenotypic measures is available from the baseline study of the Southall And Brent REvisited (SABRE) population-based cohort. SABRE was initiated in the late 1980s in North-west London with the aim of studying ethnic differences in cardiovascular disease and diabetes. The study includes people of European, South-Asian and African-Caribbean descent, aged 40–69 years at baseline [4].

***Corresponding author: Marco Molinari**, UCL, Statistical Science, London, UK, E-mail: marco.molinari.16@ucl.ac.uk. https://orcid.org/0000-0002-3374-9099
Maria de Iorio, Yale-NUS College, Singapore, Singapore, E-mail: m.deiorio@ucl.ac.uk
Nishi Chaturvedi, Alun Hughes and Therese Tillin, UCL, Population Science & Experimental Medicine, London, UK, E-mail: n.chaturvedi@ucl.ac.uk (N. Chaturvedi), alun.hughes@ucl.ac.uk (A. Hughes), t.tillin@ucl.ac.uk (T. Tillin)

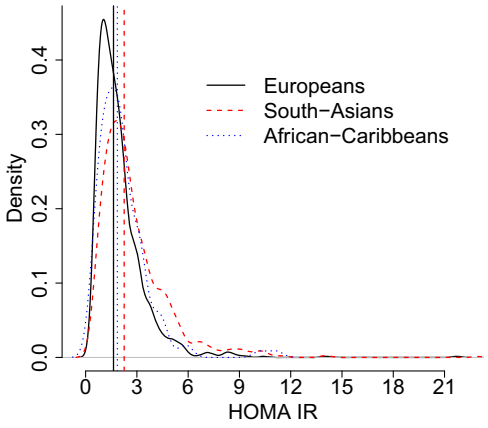

Figure 1: HOMA IR: empirical distribution by ethnicity (the vertical lines correspond to the sample median).

By employing Bayesian nonparametric statistical methods, we cluster individuals based on their HOMA IR levels. In doing so, we are able to account for the effect of covariates, in our case anthropometric measures and metabolites concentrations, identify the most influential variables and determine if the effects of covariates vary by ethnicity. We allow for clusters of individuals belonging to different ethnic groups.

Measurements of over 200 metabolites or ratios of metabolites, obtained through nuclear magnetic resonance spectroscopy, are available for more than 3000 stored baseline serum samples [5]. Lipoproteins are classified according to their density (very-low-density lipoprotein [VLDL], low-density lipoprotein [LDL], intermediate-density lipoprotein [IDL] and high-density lipoprotein [HDL]). Each lipoprotein subclass can be further characterised by its lipid composition (i.e. triglycerides, phospholipids, free cholesterol and cholesterol esters) and its particle size. The full list of metabolites included in the analysis is available in Table 1 in Supplementary Material. We include three important enzymes, alanine aminotransferase, aspartate aminotransferase and gamma glutamyl transferase. Anthropometric variables are also included, in particular global measures of body fat distribution such as waist to hip ratio (WHR) and more specific adiposity measures, such as sagittal diameter and subscapular skinfold thickness. The full list of anthropometric and clinic covariates can be found in Table 2 in Supplementary Material. We exclude from the analysis individuals with known diabetes since they were already receiving anti-diabetes medication or had undergone lifestyle modifications that might alter their metabolite levels and potentially the conclusions of the analysis. In this paper, we focus on the SABRE study baseline metabolic and phenotypic dataset. To address our research aims, we use a Bayesian nonparametric prior, the Dependent Generalized Dirichlet Process (DGDP) [6], within a regression framework. The discrete nature of the DGDP allows for data-driven clustering of the observations. We specify the DGDP prior on the regression intercept and the error precision parameter, allowing for cluster specific locations and precisions. The choice of the DGDP allows a great flexibility, accounts for inter-subject variability and it does not fix a priori the number of clusters. When prior evidence is available, through the calibration of the DGDP hyper-parameters, we can favour a large number of clusters, allowing estimation of more heterogeneous groups. Moreover, to deal with the large number of clinical and anthropometric covariates and metabolites available, we adopt a Spike and Slab approach [7, 8] in order to perform variable selection on the design matrix and highlight the most important determinant of the clinical outcome under study.

The paper is organised as follows. Section 2 introduces the statistical model. In Section 3 we present the results of the analysis and discuss the relevance of such results from the clinical point of view. Section 4 concludes the paper with a discussion and summary on the main achievements of this work.

2 The model

Let $y = (y_1, \ldots, y_n)$ be a variable observed over n individuals. We assume a linear regression model:

$$y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij} + \varepsilon_i \tag{1}$$

where p is the number of independent variables (including the intercept). The error terms ε_i are assumed to be normally distributed as

$$\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \tau^2)$$

with mean 0 and precision τ^2. The model in Eq. (1) assumes the same parameters for each observation. This assumption can be relaxed by allowing, for example, β_0 and τ^2 to vary with i (random effect model), accounting for inter-subject variability:

$$y_i = \beta_{i0} + \sum_{j=1}^{p-1} \beta_j x_{ij} + \varepsilon_i \tag{2}$$

where

$$\varepsilon_i \overset{ind}{\sim} \mathrm{N}(0, \tau_i^2) \tag{3}$$

In this way, a subject-specific intercept and precision are introduced in the model, allowing for more flexibility. We now need to specify a prior on the model parameters. In particular, we need to choose a random effect distribution for (β_{0i}, τ_i^2). A traditional and computationally convenient choice is a Normal random effects model for β_{i0} and a Gamma distribution for τ_i^2. Instead, we prefer to opt for a nonparametric random effects distribution as, often, the parametric assumptions are too restrictive in applications. The random effects distribution needs to accommodate the heterogeneity in the population and to allow for outliers, clustering and over-dispersion. At the same time, the model should not be overly complex and should still allow computationally efficient implementation of full posterior inference. Ideally the model should be a natural generalization of a traditional random effects distribution. In the next section we describe our choice of prior distribution.

2.1 Prior distributions

The model in Eq. (2) requires the specification of a prior distribution for the vector of regression coefficients $\beta = (\beta_1, \ldots, \beta_{p-1})$, the intercept β_{i0} and the precision parameter τ_i^2. We adopt a nonparametric prior, the DGDP prior, on the vector (β_{i0}, τ_i^2). As explained below, this choice of prior distribution allows to cluster the observations. Moreover the use of the DGDP prior provides both flexibility and parsimony about the number of parameters that we introduce in the model. We now present a brief review of the main properties of the Dirichlet Process (DP) and its generalisation to the DGDP. The DP (see Ref. [9, 10] for basic properties) is arguably the most widely used nonparametric Bayesian prior, mainly because of computational simplicity and ease of interpretation. In DP based models computational complexity of posterior simulation is in theory dimension independent, allowing specification of possibly high dimensional random effects distributions. A random measure P that is generated by a DP is almost surely discrete. Sethuraman [11] provides a constructive definition of this process, showing that if a random probability measure P is distributed according to a DP, with mass parameter α and base measure G_0, then it admits the following representation:

$$P = \sum_{k=1}^{\infty} \psi_k \delta_{\theta_k} \tag{4}$$

where the atoms $\theta_1, \theta_2, \ldots$ are *iid* realisations from G_0, δ_{θ_k} is the Dirac measure that assigns mass probability one in correspondence of the location θ_k. The weights ψ_k follow a *stick-breaking* process (see Ref. [12] for details):

$$\psi_k = \phi_k \prod_{j=1}^{k-1} \left(1 - \phi_j\right), \quad k = 2, 3, \ldots \tag{5}$$

with $\phi_k \overset{iid}{\sim} \text{Beta}\,(1, \alpha)$ and $\psi_1 = \phi_1$. By construction we have $0 \leq \psi_k \leq 1$ and $\sum_{k=1}^{\infty} \psi_k = 1$. The discreteness of the DP induces clustering of the subjects in the sample based on the unique values of the random effects parameters (in our case $\theta_k = (\beta_{0k}, \tau_k^2)$), where the number K of clusters is unknown and learned from the data.

In this paper we are interested in modelling the distribution of HOMA IR in each of the three ethnic groups (i) allowing for borrowing information across groups (ii) highlighting differences and similarities (iii) accounting for the effect of covariates. To this end, we employ a generalisation of the DP proposed by Ishwaran and James [12] and Hjort [13]; the Generalised Dirichlet Process (GDP). The GDP employees a richer parametrisation in the stick-breaking construction, allowing greater flexibility in the moments of the random distributions. Consider the stick-breaking in Eq. (5), where the elements ϕ_k are draws from a Beta (1, α). In the generalisation proposed by Hjort [13], the ϕ_k still draws from a Beta distribution, but the first hyper-parameter does not need to be fixed to one. In what follows we use an alternative parametrisation of the Beta distribution, where the hyper-parameters are specified in terms of the mean and the concentration parameter. In the GDP the $\{\phi_k\}$ are then independent draws from a $\text{Beta}\,(\mu_k \upsilon_k, (1 - \mu_k)\upsilon_k)$:

$$p\left(\phi_k \middle| \upsilon_k, \mu_k\right) = \frac{\Gamma\left(\upsilon_k\right)}{\Gamma\left(\upsilon_k \mu_k\right)\Gamma\left(\upsilon_k\left(1 - \mu_k\right)\right)} \phi_k^{\upsilon_k \mu_k - 1} \left(1 - \phi_k\right)^{\upsilon_k\left(1 - \mu_k\right) - 1} \tag{6}$$

where

$$\text{E}\left(\phi_k\right) = \mu_k \in (0, 1)$$

and

$$\text{Var}\left(\phi_k\right) = \frac{\mu_k\left(1 - \mu_k\right)}{\left(1 + \upsilon_k\right)}$$

with $\upsilon_k \in (0, \infty)$, are the expected value and the variance of the Beta random variable respectively. The weights of the GDP admit the same stick-breaking construction as for the DP. Hjort [13] proposes a more parsimonious parametrisation of Eq. (6), setting $\mu_k = \mu$ and $\upsilon_k = \upsilon$. This simplification does not impose significant restriction in applications. We now explain how we introduce ethnicity information in the distribution of HOMA IR. The final model will contain two main components: one for the clinical covariates and one for the patients effect. The model for the covariates expresses prior information on how covariate influence the clinical outcome, while the nonparametric prior (GDP) is used as random effect distribution to capture inter-patients variability. Moreover, it is desirable to specify a random effect distribution for each ethnicity in a way that the random effect distributions are related (similar or very different), but not necessarily identical.

There is a wealth of literature on how to extend the DP to incorporate covariate information, for example, letting the weights and/or locations of the infinite mixture in Eq. (4) depend on a variable of interest that defines sub-groups in the observations. See the seminal paper of MacEachern [14] on the Dependent Dirichlet Process (DDP). Similarly, also the GDP can be extended in presence of categorical covariates. Barcella et al. [6] introduces the DGDP, where the dependence among random distributions is introduced through the weights ψ_k of the mixture in four. The parameters ψ_k are generated from the stick-breaking process, so the dependence is introduced directly on the parameters υ and μ.

Consider G groups defined by a covariate of interest $g \in \mathcal{G}$, where $\mathcal{G}$ is the covariate space. We let μ, which represents the mean of the Beta random variables depend on the particular value of g, while we assume the same υ across groups. We denote with μ_g the mean of the Beta random variable corresponding to group g. In our application groups are defined by the ethnicity. The random measure P_g, i.e. the random distribution associated to group g, is then defined as:

$$P_g = \sum_{k=1}^{\infty} \psi_{k,g} \delta_{\theta_k}$$

Here $\theta_k = (\beta_{0k}, \tau_k^2)$. In particular, dependence across the $\mu_g, g = 1, \ldots, G$, is obtained by specifying a Beta regression on μ_g and using a categorical predictor $\mathbf{z}_g$, i.e. an indicator variable which denotes to which group the observations are associated to. This strategy allows for group dependent clustering of the observations. See Ref. [6] for details and clustering properties. Finally, the full model for HOMA IR is specified as follows

$$\begin{aligned}
y_{1g}, \ldots, y_{ng} | g, P_g &\overset{ind}{\sim} \int \mathrm{N}\left(\beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j, \tau^2\right) P_g\left(d\beta_0 d\tau^2\right) \\
P_1, \ldots, P_G \Big| \upsilon, \mu_g, G_0 &\sim \mathrm{DGDP}\left(\upsilon, \mu_g, G_0\right) \\
G_0\left(\beta_0, \tau^2\right) | m_0, \kappa_0^2, \tau_a, \tau_b &= \mathrm{N}\left(m_0, \kappa_0^2\right) \times \mathrm{Gamma}\left(\tau_a, \tau_b\right) \\
f\left(\mu_g\right) &= \mathbf{z}_g \boldsymbol{\eta} \\
\upsilon | a_\upsilon, b_\upsilon &\sim \mathrm{Gamma}\left(a_\upsilon, b_\upsilon\right)
\end{aligned}$$

for $g = 1, \ldots, G$. The parameter μ is linked through a function f (e.g. logit or probit), mapping from $(0,1)$ into $(-\infty, \infty)$, to the linear predictor $\mathbf{z}_g\boldsymbol{\eta}$ where $\boldsymbol{\eta}$ is a vector of regression coefficients of appropriate dimension to which we assign a standard Normal prior:

$$\boldsymbol{\eta} \sim \mathrm{N}\left(\boldsymbol{\eta}_\mu, \boldsymbol{\eta}_\Sigma\right)$$

where $\boldsymbol{\eta}_\mu$ and $\boldsymbol{\eta}_\Sigma$ denote the prior mean and covariance matrix respectively. We assume independence a priori between the parameters β_0 and τ^2, which is reflected in the choice of the base measure G_0, defined as the product of a Normal distribution and a Gamma distribution.

We specify a Spike and Slab prior on each of the $p-1$ regression coefficients β_j. This prior specification provides an effective variable selection strategy [7, 15]. We introduce indicator variables ω_j:

$$\beta_j = \omega_j \mathrm{N}\left(\mu_\beta, \tau_\beta^2\right) + (1 - \omega_j)\delta_0\left(\beta_j\right) \tag{7}$$

where $p(\omega_j = 1 | \pi) = \pi$ is the probability of the slab, i.e. the probability that a covariate is included in the model, while $1 - \pi$ represents the probability of the spike, i.e. the probability that the regression coefficient corresponding to the jth covariate is equal to 0 and does not affect the response. As before, $\delta_0(\beta_j)$ is a mass point at zero, representing the spike of the mixture. μ_β represents the prior mean (usually set to 0) of the slab component and τ_β^2 is the prior precision.

The parameter π is assigned a Beta prior:

$$\pi \sim \mathrm{Beta}\left(\pi_a, \pi_b\right) \tag{8}$$

where π_a and π_b are the hyper-parameters of the Beta distribution (e.g. setting $\pi_a = \pi_b = 1$ gives a uniform distribution). Appropriate choices of these hyper-parameter allows us to impose sparsity in the variable selection.

2.2 Posterior inference

Posterior inference is performed through Markov Chain Monte Carlo (MCMC) methods. A detailed description of the algorithm is provided in Supplementary Material. We run the MCMC for 30,000 iterations, discarding a burn-in period of 10,000, thinning every five iterations. We specify the following hyper-parameters: the truncation level of the stick-breaking is set to $L = 30$. This threshold works well in our application since the number of non-empty clusters is always below 20. The base measure parameters are set to $m_0 = 0$, $\kappa_0^2 = 0.1$, $\tau_a = 0.5$, $\tau_b = 0.5$. The DGDP concentration parameter υ has a

Gamma prior with $a_\upsilon = 2$ and $a_\upsilon = 1$, with expected value $a_\upsilon/b_\upsilon = 2$. The regression coefficient $\boldsymbol{\eta}$ is parametrised with prior mean $\boldsymbol{\eta}_\mu = (0, 0, 0)$ and $\boldsymbol{\eta}_\Sigma = \boldsymbol{I}_3$. The slab of the regression coefficient β_j is a Normal distribution with prior mean $\mu_\beta = 0$ and prior precision $\tau^2_\beta = 0.1$. The prior inclusion probability π has a Beta prior with parameters $\pi_a = \pi_b = 1$.

3 Results

We employ the proposed model to analysis data from the SABRE study. The empirical distribution of the outcome of interest, the Homeostatic Model Assessment insulin resistance (HOMA IR) is shown for the three ethnic groups in Figure 1. Particularly noticeable is the difference between the distribution of HOMA IR in Europeans and South-Asians. The South-Asian distribution is slightly shifted to the right and has a heavier right tail, indicating a higher percentage of more insulin resistant individuals. The distribution shows multiple local modes, pointing towards the existence of multiple sub-populations in the sample.

3.1 HOMA IR: cluster analysis

Posterior inference for HOMA IR shows evidence of 10 clusters. We use the Binder loss function [16] available in the R package *mcclust*, to estimate the number of clusters in the sample and the clustering allocation based on the MCMC output. In Figure 2 we show the empirical distribution of the outcome HOMA IR in each of the 10 estimated clusters. The overlap between some of the curves is due to the fact that the clusters are estimated conditionally to the covariates and metabolite levels included in the regression model. Table 1 summarises the ethnic composition of each cluster, while Table 2 provides some basic information in terms of age, smoking habits, percentage of females and percentage of first generations (i.e. foreign-born) migrants in each cluster. It is worth noting that clusters 8, 9 and 10, the most insulin resistant clusters, are mostly composed of first generation migrants. In Table 3 we report the ethnic composition of each cluster based on the sub-region of origin. The majority of South-Asians come from the Punjabi-Sikh minority, which represents the major South-Asian component in each cluster, with the exception of cluster 9, where there is a higher percentage of South-Asians of Muslim origin. To understand which covariates are the most important determinant of the response, we examine the posterior probability of each regression coefficient to be different from zero,

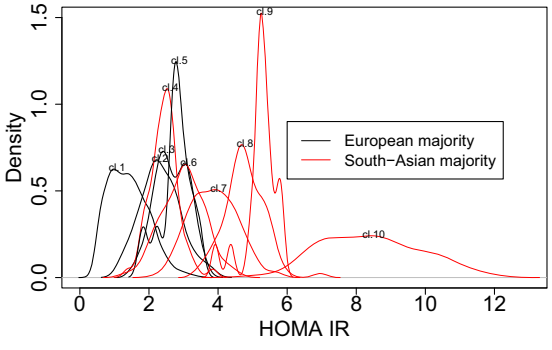

Figure 2: Empirical distribution of HOMA IR in each cluster. Black lines denote clusters with a higher proportion of Europeans, while red lines denote a higher proportion of South-Asians. A description of cluster main characteristics is given in Table 1. The numbers above each distribution denote the cluster.

Table 1: Cluster specific ethnic membership.

Cluster number	Europeans	South-Asians	Africans-Caribbean	Total number
1	784	382	83	1249
2	33	0	0	33
3	49	0	0	49
4	0	61	0	61
5	10	0	0	10
6	90	125	28	243
7	60	104	4	168
8	13	75	0	88
9	0	19	0	19
10	13	49	2	64

Table 2: Mean age, proportion of smoke habits, proportion of females and proportion of first generation migrants in each cluster. Sample standard deviations in parentheses.

Cluster number	Mean age		Ex-smoker		Current smoker		Females proportion		First generation	
1	52.51	(0.21)	0.27	(0.01)	0.25	(0.01)	0.18	(0.01)	0.46	(0.01)
2	53.27	(1.29)	0.33	(0.08)	0.45	(0.09)	0.03	(0.03)	0.21	(0.07)
3	52.02	(1.03)	0.41	(0.07)	0.31	(0.07)	0.10	(0.04)	0.10	(0.04)
4	49.54	(0.92)	0.11	(0.04)	0.16	(0.05)	0.33	(0.06)	0.98	(0.02)
5	52.00	(1.93)	0.20	(0.13)	0.30	(0.14)	0.00	(0.00)	0.10	(0.09)
6	51.93	(0.45)	0.23	(0.03)	0.19	(0.03)	0.12	(0.02)	0.67	(0.03)
7	51.43	(0.52)	0.17	(0.03)	0.19	(0.03)	0.11	(0.02)	0.67	(0.04)
8	51.25	(0.70)	0.11	(0.03)	0.12	(0.04)	0.09	(0.03)	0.86	(0.04)
9	51.89	(1.53)	0.11	(0.07)	0.37	(0.11)	0.11	(0.07)	0.95	(0.05)
10	50.95	(0.90)	0.17	(0.05)	0.25	(0.05)	0.09	(0.04)	0.80	(0.05)

Table 3: Proportion of Ethnic sub-groups of origin in each cluster (Max per row in underlined).

Cluster number	Africans Caribbean	Gujarati Hindu	Irish	Muslim	Native British	Other Europeans	Other South-Asians	Punjabi Hindu	Punjabi Sikh
1	0.07	0.02	0.07	0.04	<u>0.53</u>	0.03	0.04	0.04	0.16
2	0.00	0.00	0.18	0.00	<u>0.73</u>	0.09	0.00	0.00	0.00
3	0.00	0.00	0.10	0.00	<u>0.84</u>	0.06	0.00	0.00	0.00
4	0.00	0.16	0.00	0.13	0.00	0.00	0.13	0.11	<u>0.46</u>
5	0.00	0.00	0.00	0.00	<u>0.90</u>	0.10	0.00	0.00	0.00
6	0.12	0.03	0.04	0.08	<u>0.32</u>	0.01	0.07	0.04	0.29
7	0.02	0.07	0.01	0.11	<u>0.33</u>	0.02	0.06	0.05	<u>0.33</u>
8	0.00	0.08	0.03	0.10	0.11	0.00	0.12	0.11	<u>0.43</u>
9	0.00	0.21	0.00	<u>0.26</u>	0.00	0.00	0.21	0.16	0.16
10	0.03	0.11	0.00	0.16	0.19	0.02	0.08	0.03	<u>0.39</u>

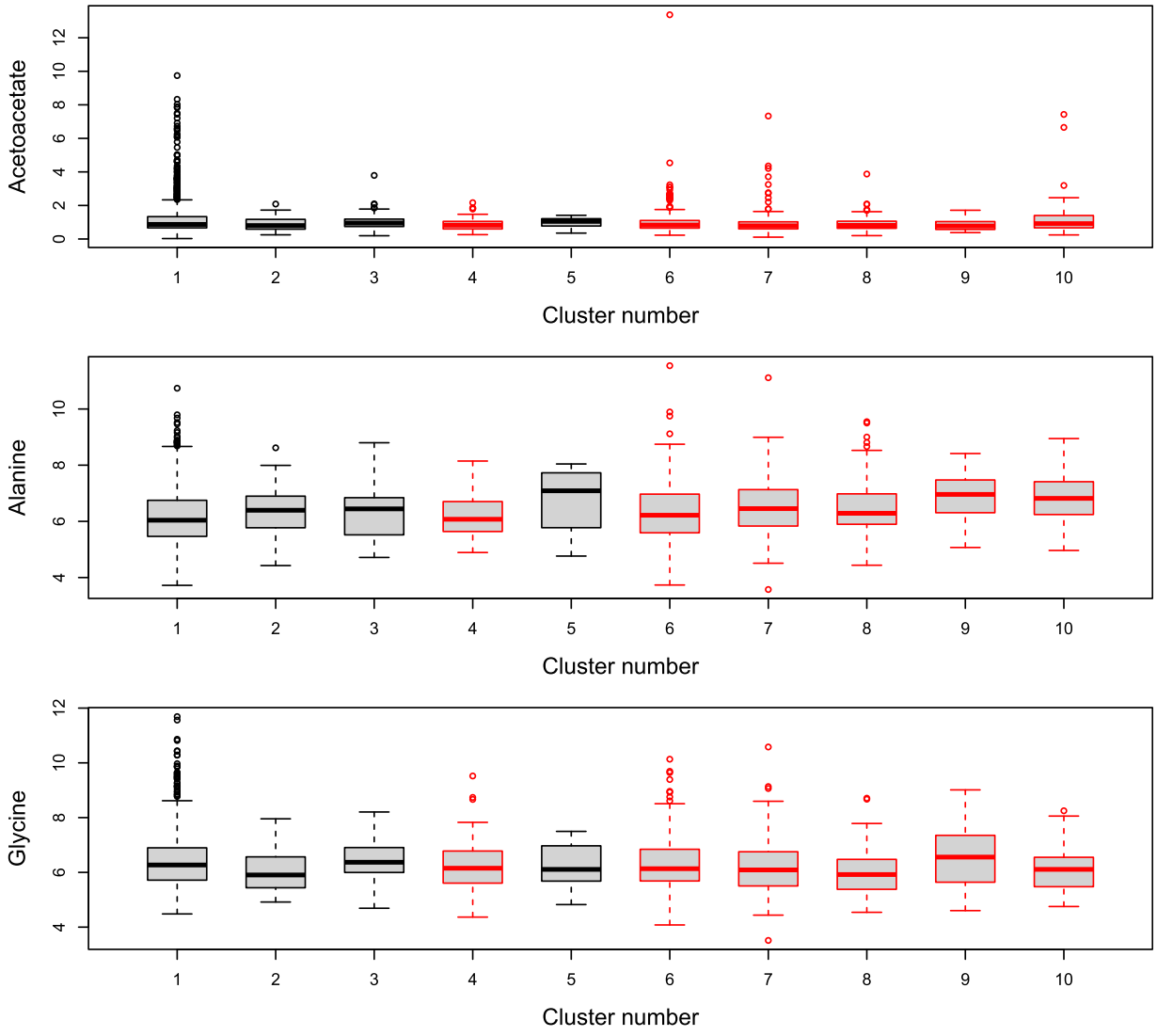

Figure 3: From top to bottom, boxplots of Acetoacetate, Alanine and Glycine. Black boxplots indicate clusters with a majority of Europeans, while red boxplots indicate clusters with a majority of South-Asians.

$p(\omega_j = 1|rest, data)$. 10 predictors have the respective $p(\omega_j = 1|rest, data) > 0.5$ and are considered for further analysis (Figures 3–5).

Cluster 1 is the largest and least insulin resistant group (n = 1249, 57% of participants). Its ethnic composition is: 71% of Europeans, 39% of South-Asians and 70% of Africans-Caribbean. The second largest group is cluster 6, comprising 243 participants (11% of the total, of which, 8% of Europeans, 13% of South-Asians and 24% of Africans-Caribbean). It is evident the net distinction between clusters with a South-Asian majority, compared with Europeans, which are all characterised by higher levels of HOMA IR, with the exception of cluster 4. Cluster 4 is entirely composed of South-Asians. In particular the cluster is characterized by almost only first-generation migrants, a higher proportion of females (0.33), relative to the other clusters

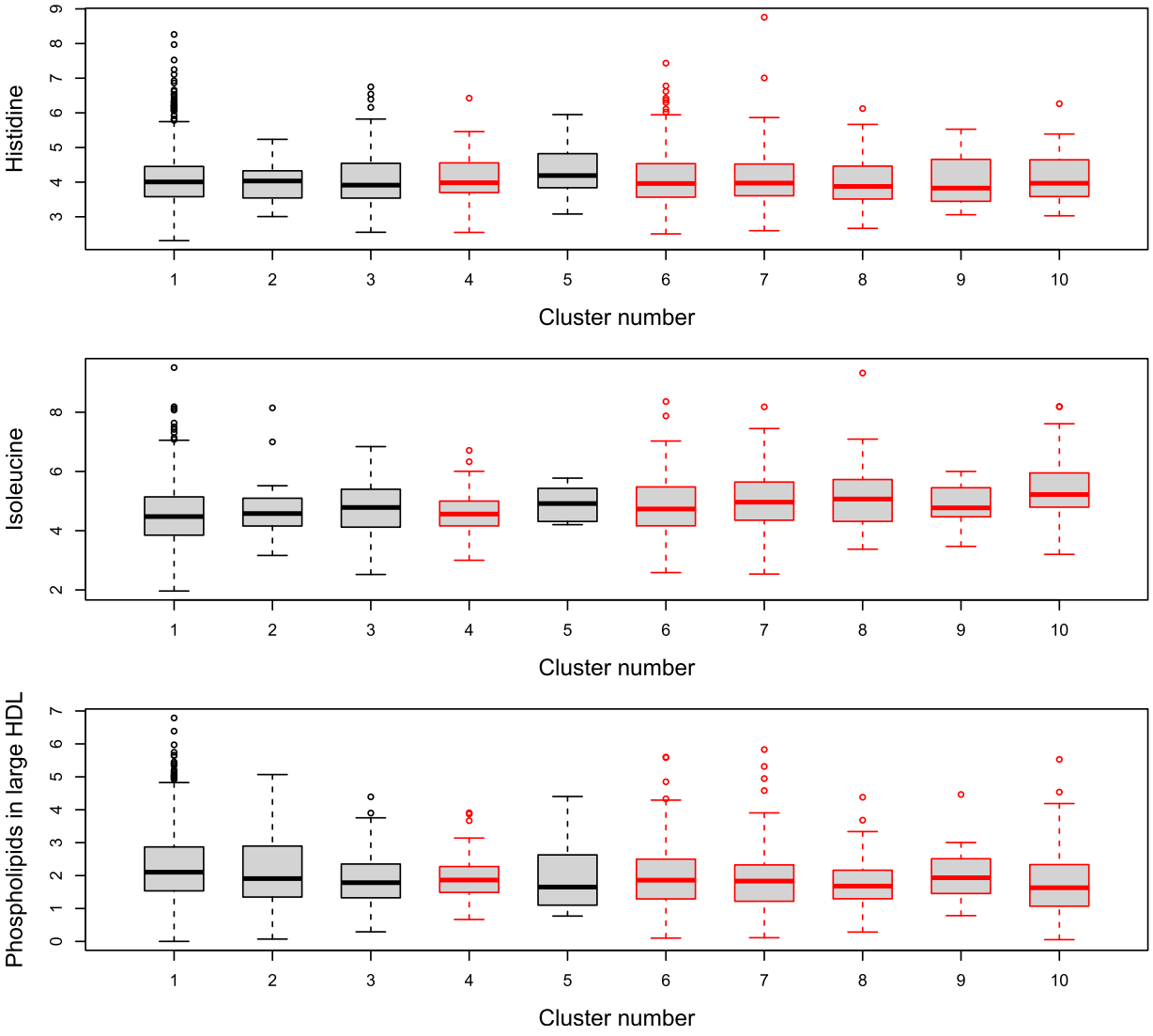

Figure 4: From top to bottom, boxplots of Histidine, Isoleucine and Phospholipids in large HDL. Black boxplots indicate clusters with a majority of Europeans, while red boxplots indicate clusters with a majority of South-Asians.

and a lower proportion of current smokers (0.16), compared with the other cluster containing South-Asians. Cluster 5 (entirely Europeans) compared with cluster 1, presents both measures of adiposity(subscapular skinfold and sagittal diameter) are modestly higher, while levels of the amino acids tyrosine and isoleucine are significantly higher. Moreover, acetoacetate levels are lower compared to cluster 1, while the levels of alanine aminotransferase (ALT) in cluster 5 are higher than in cluster 1 (Figure 5), the latter suggesting that raised HOMA IR levels may be characterised in this cluster by increased insulin levels with reduced clearance of insulin by the liver [17], implying relatively intact pancreatic beta cell function. The metabolite patterns for cluster 5 also indicate associations with both central and subcutaneous adiposity and amino acid perturbations.

Each of the 10 clusters has a distinctive metabolic and phenotypic profile, consistent with suggestions that there are different pathways to type 2 diabetes [18] and that some pathways may be more strongly associated with a particular ethnic group. For example clusters 4 and 9 are entirely composed of South-Asians, while clusters 2, 3 and 5 are entirely Europeans. Of these clusters, 8, 9 and 10 are among the most

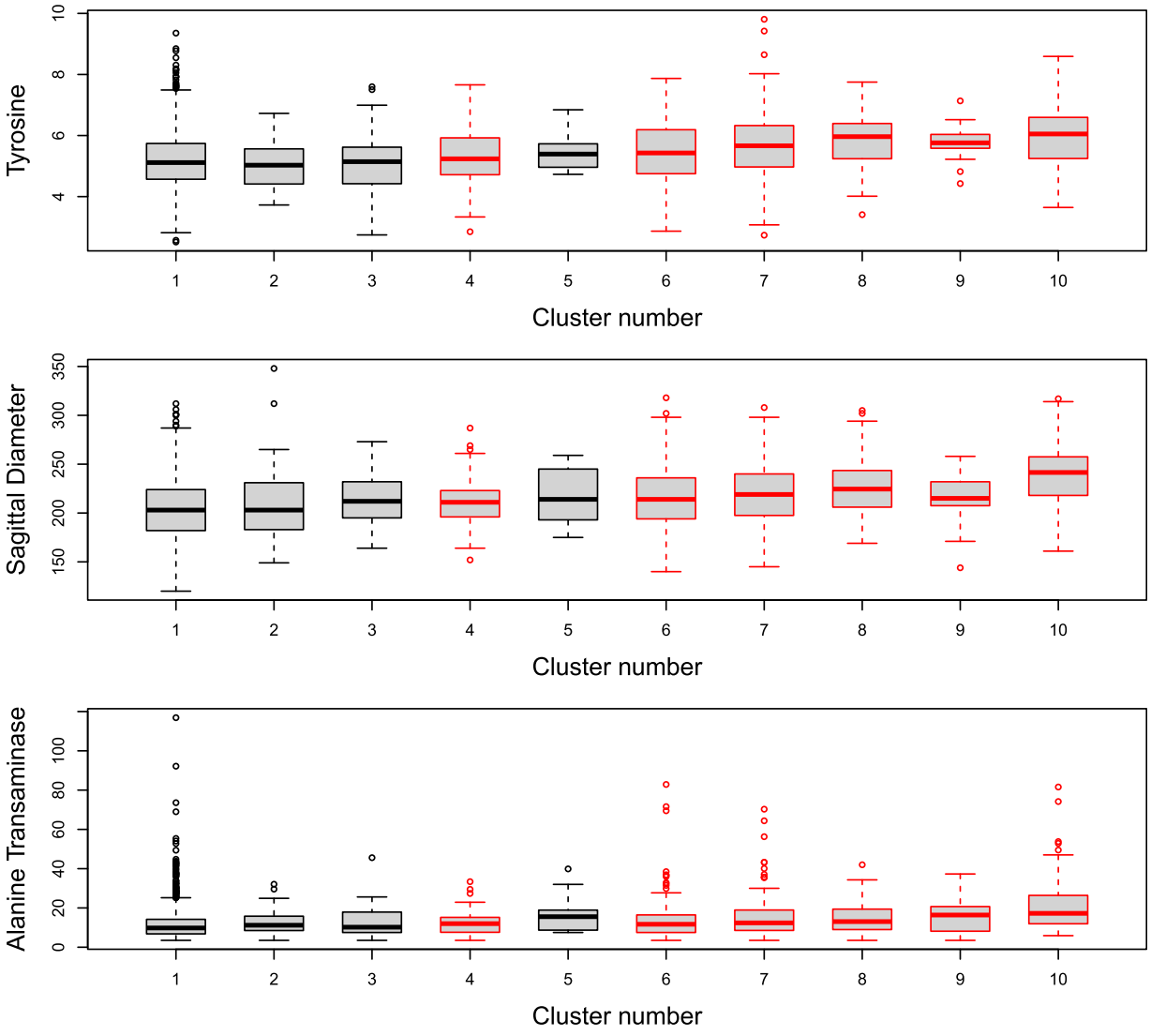

Figure 5: From top to bottom, boxplots of Tyrosine, sagittal diameter and Alanine Aminotransferase. Black boxplots indicate clusters with a majority of Europeans, while red boxplots indicate clusters with a majority of South-Asians.

insulin resistant with high levels of tyrosine, alanine, ALT and subcutaneous adiposity. Some of the clusters identified are very small and will need replication in larger studies together with formal pathway analysis. However, these methods have generated intriguing, novel and persuasive clusters, which highlight the complexity and potential multiplicity of mechanisms underlying the development of insulin resistance and type 2 diabetes.

4 Conclusions

This paper proposes the use of a nonparametric random intercept model, through the adoption of a Dependent GDP prior on the intercept coefficient and precision parameter of a linear regression. Alternative nonparametric Bayesian priors, such as the Hierarchical Dirichlet Process [19] or the Probit Srick-breaking Process [20]. The DGDP allows us to analyse multiple groups of patients and provides data-driven clustering of the

observations thanks to the Bayesian nonparametric prior. The random probability measures share the same sets of atoms, but the weights associated to each atoms differ (slightly or abruptly) across measures P_g, i.e. across groups. This construction is extremely flexible as it covers a large class of distributions. Moreover, varying weights can provide probability measures which can be remarkably close (when weights are similar), as well as probability measures which are far apart (when the weights are dissimilar). See, for example, Hatjispyros et al. [21].

We specify a spike and slab prior on the regression coefficients to effectively performs variable selection on the covariates, allowing us to understand which variables are more important in predicting the dependent variable of interest, i.e. HOMA IR. We employ the proposed model to analyse the data from the SABRE cohort study, a tri-ethnic information rich dataset on cardiovascular and metabolic diseases. Our clinical interest focuses on modelling the distribution of HOMA IR. We include anthropometric variables and metabolites concentrations as covariates in the regression framework. The results highlight the presence of sub-populations in the data, with a multi-ethnic composition, characterised by different levels of HOMA IR, which can lead to a different risk of developing type 2 diabetes. From the analysis, it is evident that cluster with higher levels of insulin resistance are composed mainly by the South-Asian ethnicity and, in particular, the more extreme clusters present a higher proportion of first-generation migrants. The results obtained from our analysis are promising and the proposed model has the potential to highlight areas for further research.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: SABRE was funded at baseline by the UK Medical Research Council and Diabetes UK. Follow-up studies have been funded by the Wellcome Trust (WT 082464), British Heart Foundation (SP/07/001/23603 and CS/13/1/30327) and Diabetes UK (Metabolomics: 13/0004774). Nishi Chaturvedi received support from the National Institute for Health Research University College London Hospitals Biomedical Research Centre.
Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

1. Sproston K, Mindell J. Health survey for England 2004. The health of minority ethnic groups. Leeds: The Information Centre; 2006.
2. Tillin T, Hughes AD, Godsland IF, Whincup P, Forouhi NG, Welsh P, et al. Insulin resistance and truncal obesity as important determinants of the greater incidence of diabetes in indian asians and african caribbeans compared with europeans: the southall and brent revisited (SABRE) cohort. Diabetes Care 2013;36:383–393.
3. Matthews DR, Hosker JP, Rudenski AS, Naylor BA, Treacher DF, Turner RC. Homeostasis model assessment: insulin resistance and β-cell function from fasting plasma glucose and insulin concentration in man. Diabetologia 1985;28:412–9.
4. Tillin T, Forouhi NG, McKeigue PM, Chaturvedi N. Southall and brent revisited: cohort profile of sabre, a UK population-based comparison of cardiovascular disease and diabetes in people of european, indian asian and african caribbean origins. Int J Epidemiol 2010;41:33–42.
5. Soininen P, Kangas AJ, Würtz P, Suna T, Ala-Korpela M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. Circ: Cardiovasc Genet 2015;8:192–206.
6. Barcella W, De Iorio M, Favaro S, Rosner GL. Dependent generalized Dirichlet process priors for the analysis of acute lymphoblastic leukemia. Biostatistics 2017;19:342–58.
7. George EI, McCullogh RE. Variable selection via Gibbs sampling. J Am Stat Assoc 1993;88:881–9.
8. George EI, McCulloch RE. Approaches for Bayesian variable selection. Stat Sin 1997;7:339–73.
9. Ferguson TS. A Bayesian analysis of some nonparametric problems. Ann Math Stat 1973;1:209–30.
10. Antoniak CE. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann Math Stat 1974;2:1152–74.
11. Sethuraman J. A constructive definition of Dirichlet priors. Stat Sin 1994;4:639–50.
12. Ishwaran H, James LF. Gibbs sampling methods for stick-breaking priors. J Am Stat Assoc 2001;96:161–73.
13. Hjort NL. Bayesian analysis for a generalised dirichlet process prior. Preprint series. Statistical Research Report; 2000. Available from: https://urn.nb.no/URN:NBN:no-23420.

14. MacEachern SN. Dependent nonparametric processes. In: ASA proceedings of the section on Bayesian statistical science. Alexandria, Virginia. Virginia: American Statistical Association; 1999: 50–5 pp.
15. Malsiner-Walli G, Wagner H. Comparing Spike and Slab Priors for Bayesian Variable Selection. Austrian J Stat 2011;40:241–64.
16. Binder DA. Bayesian cluster analysis. Biometrika 1978;65:31–8.
17. Bonnet F, Ducluzeau P-H, Gastaldelli A, Laville M, Anderwald CH, Konrad T, et al. Liver enzymes are associated with hepatic insulin resistance, insulin secretion, and glucagon concentration in healthy men and women. Diabetes 2011;60:1660–7.
18. Udler MS, Kim J, von Grotthuss M, Bonas-Guarch S, Cole JB, Chiou J, et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: a soft clustering analysis. PLoS Med 2018;15:e1002654.
19. Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. J Am Stat Assoc 2006;101:1566–81.
20. Rodriguez A, Dunson DB. Nonparametric Bayesian models through probit stick-breaking processes. Bayesian Anal 2011;6: 145–77. (Online).
21. Hatjispyros SJ, Nicoleris T, Walker SG. Random density functions with common atoms and pairwise dependence. Comput Stat Data Anal 2016;101:236–49.

Supplementary material: The online version of this article offers supplementary material (https://doi.org/10.1515/ijb-2019-0108).